JOURNAL OF YOUTH IMPACT

ISSN: 3070-6432

Volume I · Issue I 2026 January

Featured Articles

- Cooperative Integration of SCR and CESAR-1
- The Impact of Query Decomposition and Cross-Encoder Reranking
- Toward Harmonized Global Governance of GMOs
- A Media Affordance Perspective on Digital Hoarding Behaviour
- Leveraging Machine Learning for Stroke Prediction

journaljyi.org

About the Journal of Youth Impact

The Journal of Youth Impact (ISSN: 3070-6432) is a prestigious peer-reviewed academic journal dedicated to publishing high-quality research conducted by high school, undergraduate, and graduate students from around the world. Founded with the mission of bridging emerging student researchers and the global academic community, JYI provides a rigorous and authentic platform for original scholarly inquiry across disciplines.

JYI embraces research spanning the full spectrum of academic fields — from natural sciences and engineering to humanities and social sciences. We are committed to cultivating methodological rigor, intellectual curiosity, and academic integrity in every young researcher, offering a genuine academic publishing experience that mirrors professional scholarly standards.

To uphold this mission, JYI has assembled a distinguished editorial and review board of nearly 100 scholars from Ivy League universities, Oxford, Cambridge, and other leading global institutions. These experienced editors and reviewers—many of whom serve on premier academic journals—apply the same exacting standards to student work. Through this process, JYI serves not only as a publication venue, but also as an unparalleled training ground that prepares the next generation of scholars for future academic excellence.

© Copyright Information

Table of Contents

Cooperative integration of SCR and CESAR-1, focused on mechanism and performance

Dongseok Kim, Queens College, City University of New York, 65-30 Kissena Blvd, Flushing, New York, 11367, United States
* Corresponding author email: kimdongseok3063@gmail.com

Abstract

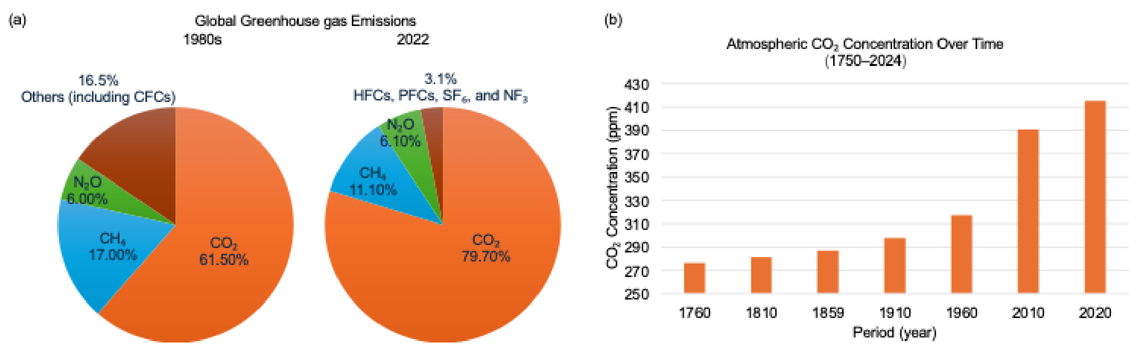
Limiting carbon dioxide (CO_2) emissions from large point sources such as thermoelectric plants is a central challenge for mitigating global warming. Among post-combustion capture methods, amine-based absorption using CESAR-1, a blended (AMP) / (PZ) solvent, has been widely used due to its higher CO_2 loading capacity and lower regeneration energy. However, the presence of nitrogen oxides (NO, NO_2) in flue gas imposes a critical limitation: during solvent regeneration, nitrogen oxides promote degradation of CESAR-1, leading to the formation of carcinogenic nitrosamines and nitramines, which are known to be toxic. This paper reviews the distinct characteristics of AMP and PZ, the two components of CESAR-1, as well as the reaction mechanism underlying CO_2 capture in CESAR-1 system. Also, it focuses on the solvent degradation pathway that occurs in the presence of nitrogen oxides, and the mechanism of selective catalytic reduction (SCR). Building on these concepts, we propose an integrated SCR-CESAR-1 system that combines SCR with CESAR-1, an amine-based solvent, to reduce CO_2 emissions while minimizing the risk of undesired byproduct formation. Furthermore, by optimizing the catalyst used in the SCR system, the suggested system aimed to minimize energy consumption across the system.

Keywords

Global warming; Carbon capture; Post-combustion carbon capture; Selective catalytic reduction; CESAR-1

Introduction

Due to rapid industrialization and rising global energy consumption, carbon dioxide (CO_2) levels in Earth's atmosphere have been increasing rapidly. CO_2 accounts for the largest share of human-generated greenhouse gases, and its concentration has continued to increase due to human activities [1], [2]. Not only had its share of greenhouse gas emissions increase over the past 250 years, but the absolute amount of CO_2 has increased as well. This increased CO_2 acts as a significant greenhouse gas, trapping the heat radiated from the Earth [3], [4]. Figure 1 illustrates the historical increase in CO_2 emissions and atmospheric concentrations, particularly since the Industrial Revolution.



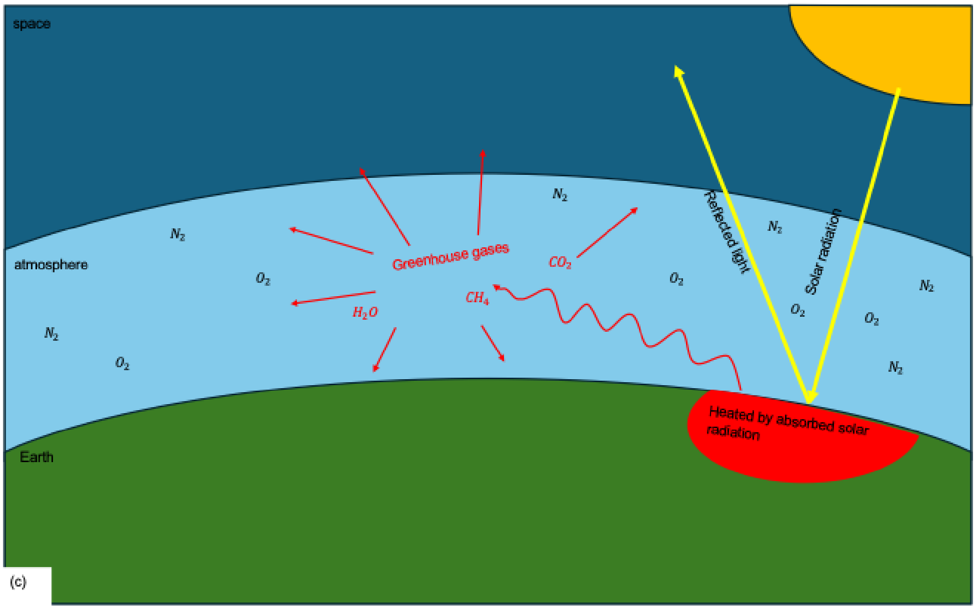


Figure 1. Increase of CO_2 after the Industrial Revolution. (a) The comparison of global greenhouse gas concentrations is shown between the 1980s and 2022. In the 1980s, CO_2 accounted for about 65% of total global greenhouse gas emissions, but in 2022, its share had increased to over 75% [1], [2]. (b) The actual amount of CO_2 continued to increase since the start of the Industrial Revolution in 1750, with a rapid increase between 1960 and 2010 [3], [4]. (c) The graphical scheme for the greenhouse effect. Greenhouse gases, including H_2O, CH_4, CO_2, etc., are atmospheric gases that absorb and emit infrared radiation particularly well, thereby significantly contributing to global warming [5]. (adopted from Schmidt et al., (2010))

Among the several technologies aimed at reducing excess atmospheric CO_2, the post-combustion method is the most well-suited and efficient way to capture and process industrial emissions, given its high compatibility with existing processes. Capture is accomplished by separating and capturing the CO_2 emitted with flue gas from industrial plants. The post-combustion method can also be adapted to the existing plant without changing the basic operating procedure (Figure 2) [6].

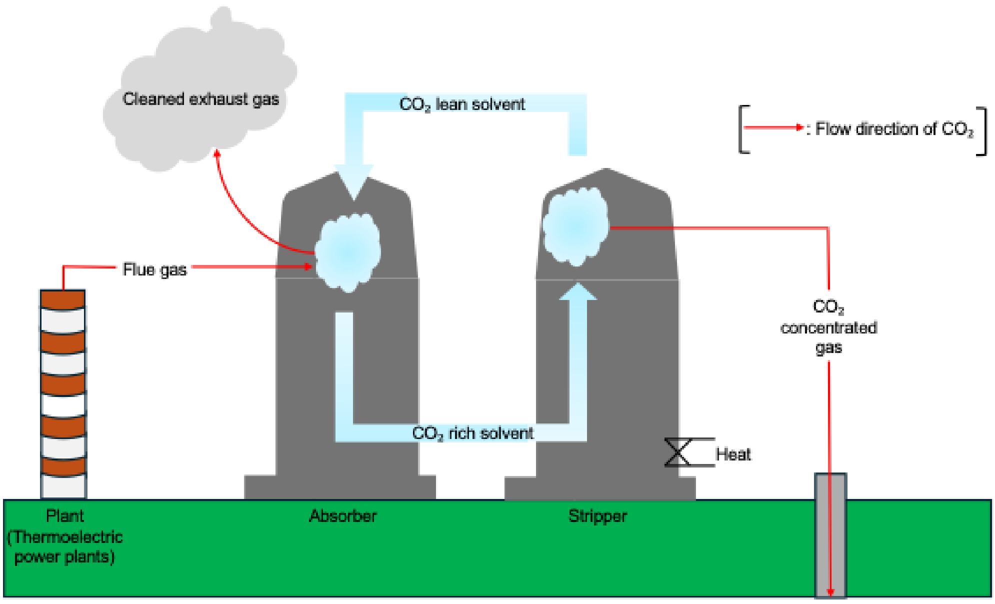
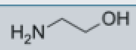
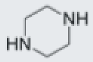


Figure 2. Structure of post-combustion system. The graphical scheme of the post-combustion system [6] (adopted from Ros et al., (2025)). When the flue gas initially enters the absorber, it reacts with the amine-based solvent, capturing carbon dioxide (CO_2). The

CO_2-depleted gas is then released into the atmosphere. Then, the CO_2-retained solvent is transported to the Stripper, and heat treatment allows the rich solvent to separate into two substances: concentrated (or highly purified) CO_2 gas and CO_2-lean solvent.

The post-combustion strategy combines an absorber with a stripper. The absorber is designed to capture CO_2 from the plant's flue gas using an amine-based solvent, and the stripper is designed to re-emit CO_2 from the CO_2-rich solvent in a highly purified state, allowing it to be used as a source of future energy resources. The post-combustion method can minimize additional equipment costs and construction time, as it does not require altering the fundamental structure. Additionally, since this system can operate independently of the combustion process that forms flue gas, post-combustion methods have been receiving increasing attention for decarbonizing thermoelectric power plants and industrial processes. Post-combustion methods use amine-based solvents, thanks to their high reactivity with CO_2 [7]. Table. 1 represents each amine-based solvent's structure, relative carbon capture efficiency, and brief descriptions of its properties. [8], [9], [10]

Table 1. Structure and categorization of MEA, MDEA, AMP, and PZ, based on functional groups [8], [9], [10]. The efficiency in carbon capture varies with the concentration of each solvent, the amine circulation rate (cubic meters/hour), and, in many cases, the mixing and utilization of the solvents.

Amine-based solvent	Structure	Comparative carbon capture efficiency	Properties/Mechanism
MEA		High	Rapid absorption rate and high solubility Form carbamate/bicarbonate
MDEA		low	Tertiary amine, Form bicarbonate, Slow absorption rate. Rare to use it alone
AMP		Medium	Primary amine, Form carbamate, excellent in terms of thermal stability/regeneration
PZ		Very high	Strong nucleophile, Form carbamate rapidly, Very fast reaction rate

Among others, MEA (Monoethanolamine) has been traditionally utilized as a primary amine-based solvent. According to Bottoms (1930), MEA was originally utilized for separating acidic gases (CO_2, SO_2, and H_2S) from gas mixtures [11]. It is a water-soluble aliphatic amino alcohol that reacts chemically with acidic gases at low temperature and releases them upon heating. It has been broadly adopted in the initial industrial field due to its relatively high reactivity with CO_2. MEA has been developed since the early industrial era, so it has sufficient technological maturity. However, the MEA simultaneously has distinct disadvantages due to its high corrosivity and requires high regeneration energy, approximately 3.6-4.2 GJ/t CO_2 [12], [13].

To address these problems, MDEA or mixtures of individual amines, such as AMP with PZ (CESAR-1), were developed [14]. One alternative, MDEA (Methyl diethanolamine), absorbs CO_2 via the bicarbonate pathway rather than carbamate formation. MDEA has high thermal stability and low regeneration energy requirements, but a slow CO_2 absorption rate. Therefore, it is rare for MDEA to be used alone. During initial plant commissioning, MDEA was chosen for its ability to remove H_2S and retain a portion of the CO_2 in the residue gas. Its greater selectivity for H_2S than for CO_2 can be helpful once H_2S is removed as an acidic gas, but to remove CO_2, it can only be captured via a bicarbonate pathway, which is a slow reaction and ultimately undesirable for its sole use in CO_2 capture [15]. Furthermore, in the presence of oxygen and metal cations, MDEA undergoes oxidative degradation, forming corrosive byproducts such as formate and peroxide.

Even single-component amine solvents, such as MEA and MDEA, still involve a trade-off between absorption performance and regeneration energy. As a result, many researchers have turned to blended amine systems, in which one amine complements the shortcomings of the other. To maximize performance in both processes, recent research has widely used a mixture of amines to compensate for each solvent's weaknesses by combining two amine-based solvents with complementary properties. One of the representative amines is the CESAR-1 solvent, or a mixture of AMP (2-amino-2-methyl-1-propanol) and PZ (piperazine).

According to Mertdağ et al. (2023), the reboiler duty, which is the amount of heat to regenerate CO_2 from solvent, of CESAR-1 is about 20% lower than MEA and 10-15% lower than MDEA/PZ, together with a low heat of absorption and reduced solvent circulation rate (L/G) [16].

Solvent circulation rate (L/G) is expressed as two factors: liquid and gas. In the CO_2 capture process, the solvent is circulated throughout the overall system, which consists of the absorber and stripper. When a certain amine solvent possesses a high CO_2 loading capacity, the whole system can circulate a smaller amount of liquid amine to capture the same amount of CO_2. As CESAR-1 has a high CO_2 loading efficiency, it requires less liquid solvent to circulate, thereby reducing the additional energy-consuming operations of pumps, heating, cooling, etc. Therefore, CESAR-1 delivers improved energy efficiency while maintaining CO_2 absorption capacity equal to or above that of other systems. According to the modeling result, CESAR-1 shows higher CO_2-rich loading. When the system absorbs 90% of CO_2 from flue gas, the AMP-PZ system needs approximately 2.7 GJ/t CO_2; on the other hand, the MDEA-PZ system needs a high energy requirement, 2.9 GJ/t CO_2, while regenerating the CO_2-rich solvent [17].

CESAR-1 has recently attracted attention as a replacement for the MEA due to its high energy efficiency and stable operation; however, it still has intrinsic limitations when NO_x is present in the flue gas, as NO_x promotes its degradation in the CESAR-1 system. In this context, x is restricted to 1 and 2 only.

Apart from NO_x influx into the system, the solvent can be degraded through two steps: thermal degradation and oxidation. NO_x, especially, affects this oxidation step, forming carcinogenic substances such as nitrosamines and nitramines even at low concentrations, 0.5 - 2.35 ppm [18]. Byproducts formed during the degradation steps have been reported to reduce overall carbon capture efficiency.

The Selective Catalytic Reduction (SCR) system is designed to resolve this issue and can utilize this flue gas, especially for NO_x. Engelhard Corporation first patented SCR technology in 1957, and since the early 1970s, SCRs have been widely used for both nitrogen oxide control from stationary combustion sources burning fossil fuels and for reducing solvent degradation in post-combustion carbon capture applications.

Taken together, CESAR-1 offers distinct advantages in energy efficiency compared to traditional MEA-based solvents; however, it also has critical drawbacks (nitrosamine/nitramine formation in a stripper process and solvent degradation), which serve as a barrier for utilization in industry sectors. This paper proposes a new flow design for the post-combustion carbon capture method, incorporating a SCR system upstream to prevent the formation of such carcinogenic substances in advance. As a review article, this paper focuses on the absorption and regeneration mechanism of CESAR-1, the degradation pathways under NO_x exposure, a SCR mechanism and associated catalyst, and proposes an integrated flow diagram of the overall process (SCR-CESAR system).

The Post-combustion Capture mechanism using CESAR-1

Mechanism in Absorber

In the CESAR-1 solvent system, CO_2 can be captured by amine species, which can be carboxylated to bicarbonate or carbamate, as shown in Figure 3.

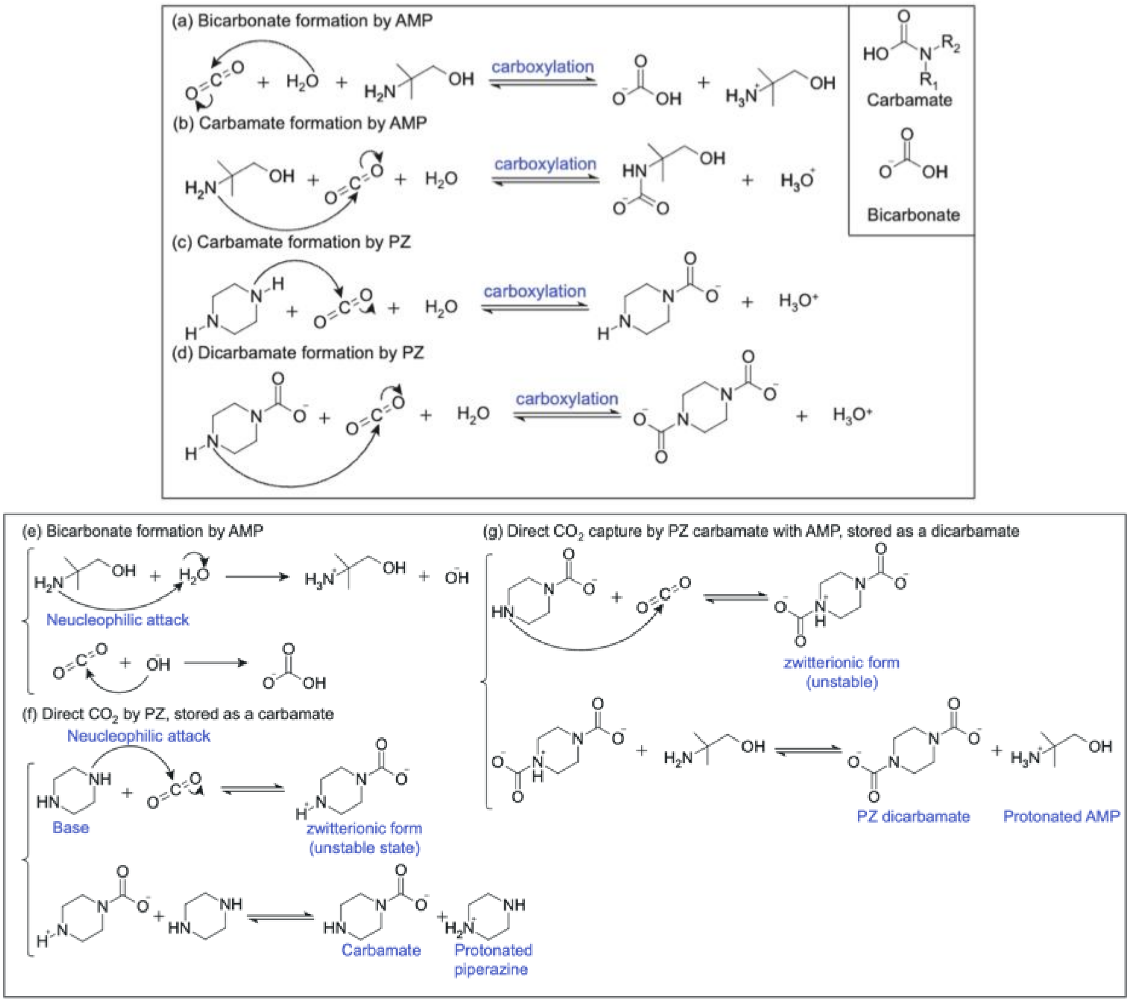


Figure 3. Carbon capture mechanism in the absorber. CO_2 is stored either in carbamate or bicarbonate form with an amine-based solvent. For (a) – (d), there are depicted brief results of the reaction mechanism in CESAR-1. Each reaction is conducted via carboxylation to capture CO_2 in the solvent. Even if AMP is still under the impression of steric hindrance by two methyl groups, a small amount of AMP-carbamate is present in stable form. For (e) – (g), more detailed processes of the reaction mechanism. Fundamental reactions involve nucleophilic attack of the lone electron pair of nitrogen on the electrophilic site, forming a protonated or zwitterionic form with charged ions [19], [20], [21].

The AMP (2-amino-2-methyl-1-propanol) and PZ (piperazine) blend acts cooperatively via two distinct mechanisms during the CO_2 capture process. AMP is mainly involved as a bicarbonate, whereas PZ tends to be reserved for the carbamate form.

For the pathway involving bicarbonate formation (Figure 3a), AMP is highly favored. The steric nature of both methyl substitutions (-CH_3) near the amine nitrogen accounts for the unfavorable nature of carbamate formation with AMP by hindering the approach of CO_2 near the Nitrogen atom. Therefore, AMP unfavorably reacts with the amine, but it does interact with water. Proton transfer equilibrium with water gives free hydroxide ions (OH^-), which attack dissolved carbon dioxide to form bicarbonate, as shown in Figure 3e [19], [20]. However, as seen in Figure 3b, the AMP-carbamate formation process is undeniably present, and it is as stable as MEA-carbamate due to a similar hydrolysis barrier, which prevents the carbamate from easily breaking apart. Nevertheless, AMP-carbamates are unlikely to be observed in the CESAR-1 system due to the kinetic preference in the bicarbonate pathway [20]. The methyl groups increase the negative charge density and polarizability of a nitrogen atom. They stabilize the binding of the hydrogen atoms in the amine group and the oxygen atom of the interaction between H_2O and AMP and destabilize the binding of CO_2. The charge redistribution also stabilizes the binding of the nitrogen atom of AMP with the surrounding H_2O molecules and makes CO_2 less accessible to the nitrogen site of the amine. Thus, the AMP-carbamate is less favorable than hydrated bicarbonate. [20]

Piperazine (PZ) reacts directly with CO_2 to form carbamate and dicarbamate species (Figures 3c–d and 3f-g). Given that there are two secondary amine centers on PZ, these nucleophilic sites react readily with CO_2, forming both carbamate and dicarbamate. This allows it to form a zwitterionic intermediate (a neutral species with both positive and negative charges), which is rapidly deprotonated to form piperazine carbamate (Figure 3f) [21]. The second amine site binds a CO_2 molecule, forming a dicarbamate (Figure 3g). These reactions indicate the direct CO_2 capture pathway, which proceeds via carboxylation and deprotonation.

Overall, the AMP/PZ combination shows a synergistic effect in CESAR-1. In particular, while AMP ensures a low energy conversion of CO_2 into bicarbonate via water-assisted hydrolysis, PZ guarantees rapid carbamate formation via direct carboxylation, improving the solvent's absorption kinetics, stability, and reducing the regeneration energy compared to conventional MEA-based solutions.

Mechanism in Stripper and degradation process

The staple regeneration mechanism is shown in Figure 4.

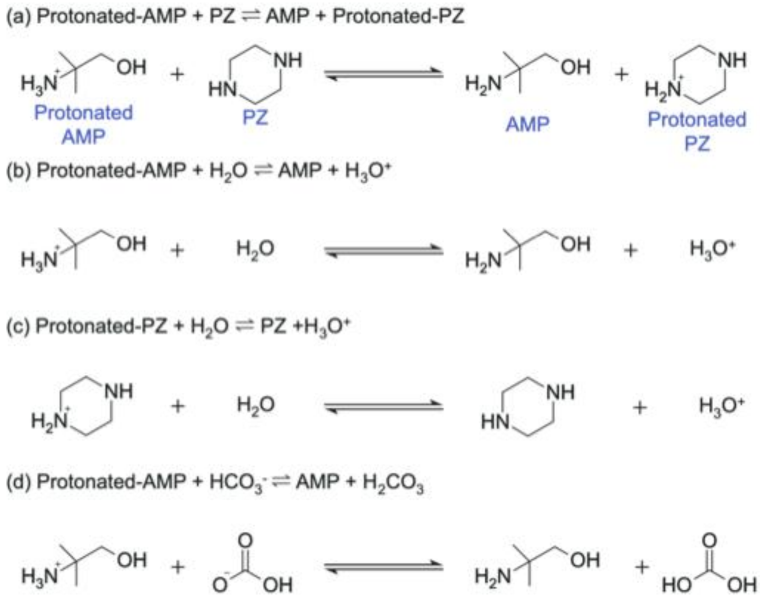


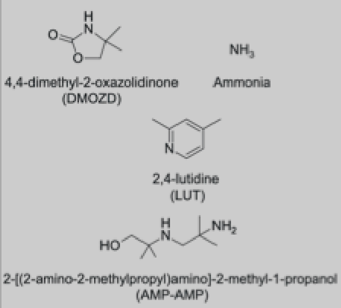
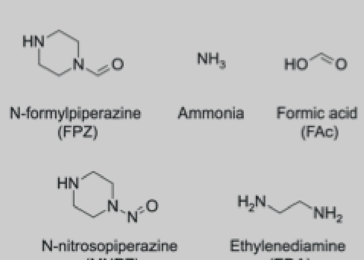
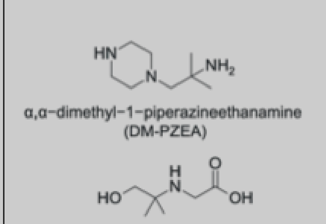
Figure 4. Regeneration mechanism in the stripper. Regeneration mechanism of each amine-based solvent (adapted from Li et al. (2023)) is described [22].

CO_2 is detached from the CO_2-rich solvent by infusing a certain range of heat into the stripper phase. This process, called regeneration, is the reverse of an absorption mechanism that restores contained CO_2 as bicarbonate in AMP and carbamate in PZ. Unlike the single-amine mechanism, the blended amine has multiple proton-transfer pathways to regenerate the original amine and produce CO_2. Furthermore, the presence of bicarbonate, which serves as a proton transfer path, is responsible for less regeneration energy than the independent use of each amine, according to Li et. al. (2023) [22]. Through these regeneration pathways, the recovered amines are transferred to the absorber and continuously undergo the same process until their life span is over.

Although the operating temperature and pressure generally vary depending on the operating conditions, the absorber operates under mild conditions around 40 ℃, 1.1 bar, where CESAR solvent can stably perform, and the stripper operates under relatively harsh conditions around 120-125 °C, 0.9 bar, which facilitates CESAR solvent regeneration [23], [24]. Since the CESAR-1 system operates over a wide temperature range and in the presence of oxygen-containing substances in the flue gas, it inevitably exhibits side effects, such as thermal degradation or oxidation, under these conditions.

The major byproducts derived from both thermal degradation and oxidation are depicted in Table 2.

Table 2. Chemical structures of byproducts generated from AMP, PZ, and AMP/PZ mixed solvent systems. The byproducts are formed during the regeneration process, in which CESAR-1 solvent unintentionally degrades into a few byproducts, taken from Buvik et al (2025) [25]. Interestingly, some degradation products do not form when AMP or PZ degrades individually but become prominent only when the two amines are mixed.

	Byproducts from AMP (Primary amine)	Biproducts from PZ (Secondary amine)	New Biproducts from CESAR-1 (AMP/PZ)
structure	 4,4-dimethyl-2-oxazolidinone (DMOZD) Ammonia (NH_3) 2,4-lutidine (LUT) 2-[(2-amino-2-methylpropyl)amino]-2-methyl-1-propanol (AMP-AMP)	 N-formylpiperazine (FPZ) Ammonia (NH_3) Formic acid (FAc) N-nitrosopiperazine (MNPZ) Ethylenediamine (EDA)	 α,α−dimethyl−1−piperazineethanamine (DM-PZEA) N-(2hydroxy-1,1-dimethylethyl)-glycine (HMeGly)

For the byproducts formed during the thermal degradation and oxidation of CESAR-1, the likelihood of returning from the CO_2-rich solvent to the lean solvent is low, which in turn reduces the CO_2 absorption capacity. Also, some organic acids, such as formic acid, acetic acid, and acetone, have been identified as major oxidation products at higher CO_2 loadings. A large amount of corrosion has been reported due to such acids produced by degradation, causing material damage and leading to equipment failure. These drawbacks from inherent thermal degradation and oxidation can serve a pivotal role in further usage of CESAR-1 in the Post-combustion carbon capture system. Researchers suggested maintaining the temperature in the stripper within a certain range of approximately 120 °C [24].

However, the presence of NO_2 in the stripper, which reacts to form N-nitrosopiperazine (MNPZ), poses a major problem due to its negative effects on the environment and humans. Specifically, increasing the risk of liver, lung, and gastrointestinal cancers upon prolonged exposure to MNPZ, and contaminating aqueous environments, was reported in Vikram et al. (2024) [26]. Furthermore, in mixtures rich in PZ, re-arrangement and hydrolysis of carbamate intermediates produce MNPZ and related nitrosamines even in the presence of low concentrations of NO_2 (30-50 ppm). These unexpected byproducts not only lead to solvent loss in the regeneration process, but also hazardous effects within or near the post carbon capture system. These impacts cannot be reduced or regulated by changing the reaction conditions, such as temperature or pressure [26]. Consequently, limiting NO_x concentrations is critical to minimize solvent loss and the associated environmental impact.

Selective Catalytic Reduction (SCR)

In SCR, NO_x is converted into harmless N_2 and H_2O with a reductant (traditionally, ammonia) by reaction on a supported catalyst's surface. The catalyst structure can reduce the associated activation energy and steer the reaction toward the N-N coupling pathway, minimizing the formation of undesired byproducts, such as NH_3 or N_2O, especially N_2O, which is considered a more potent greenhouse gas than CO_2.

In post-combustion CO_2 capture, the upstream SCR reduces the amount of NO_x entering the absorber/stripper loop, thereby reducing nitrosamines and oxidatively degrading the solvent. In general, ammonia-based SCR follows the reactions shown in Figure 5 [27].

(a) Similar concentration of NO, NO_2 (Fast SCR)

$$NO + NO_2 + 2NH_3 \rightarrow 2N_2 + 3H_2O$$

(b) NO dominant (Standard SCR)

$$4NO + 4NH_3 + O_2 \rightarrow 4N_2 + 6H_2O$$

(c) NO_2 dominant

$$6NO_2 + 9NH_3 \rightarrow 7N_2 + 12H_2O$$

Figure 5. Ammonia based SCR reaction mechanism. This set of reactions was taken from Iwasaki et al. (2010) [27] (a), Mechanism of fast SCR reaction. Fast SCR reaction becomes dominant when NO and NO_2 is at similar concentration (b), Mechanism of standard SCR reaction. Standard SCR reaction occurs when NO is dominant among NO_x substances (c), Mechanism of NO_2-dominant SCR reaction.

Based on the flue gas concentration, the tendency among the three reactions varied. According to Saito, Fujiwara (2023), the key factors in forming NO_x in a fossil fuel power plant are classified into Steam Flow Rate, Furnace Pressure, Combustion Air Temperature, and Steam Temperature near the furnace wall [28]. These factors function with either a positive or negative tendency to form NO_x. Based on these factors, the NO_x production mechanism is classified into Thermal NO_x, Fuel NO_x, and Prompt NO_x, as shown in the Figure. 6. Thermal NO_x occurred regardless of the presence of fuel and underwent a chain reaction rather than a direct reaction to form nitric oxide, because the boiler's temperature was not high enough for N-N bonding to be easily broken. On the other hand, Fuel NO_x results from the reaction of nitrogen-containing fuels with radicals, such as hydroxyl radical, oxygen radical, and hydrogen radical. At first, the nitrogen-containing fuel, in this case, ammonia, breaks one of its N-H bonds to form the NH_2 radical. This radical can help form NO by being oxidized by oxygen gas, an oxygen radical, or nitroxyl (HNO). Both Thermal NO_x and Fuel NO_x involve flame radicals, which facilitate the reactions. Prompt NO_x reaction mechanism is not considered significant, as the research stated that prompt NO_x can be negligible in the conventional combustion of pulverized coal [29], [30].

(a) Themal NO_x formation mechanism (Extended Zeldovich mechanism)

$$\begin{cases} N_2 + O \rightarrow NO + N \\ N + O_2 \rightarrow NO + O \\ N + OH \rightarrow NO + H \end{cases}$$

(b) Fuel-NO_x formation mechanism

$$\begin{cases} NH_3 + OH \rightleftharpoons NH_2 + H_2O \\ NH_3 + O \rightleftharpoons NH_2 + OH \\ NH_3 + H \rightleftharpoons NH_2 + H_2 \end{cases}$$

$$\begin{cases} NH_i + H(OH) \rightleftharpoons NH_{i-1} + H_2(H_2O),\ i = 1, 2 \\ NH + O_2 \rightleftharpoons NO + OH \\ NH + O \rightleftharpoons NO + H \end{cases}$$

$$\begin{cases} O + NH_2 \rightleftharpoons HNO + H \\ HNO + M \rightleftharpoons NO + H + M \\ HNO + OH \rightleftharpoons NO + H_2O \\ HNO + NH_2 \rightleftharpoons NO + NH_3 \end{cases}$$

M = Enhanced third-body efficiencies (H_2O, O_2, and N_2)

Figure 6. NO_x formation mechanism in power plants. There are two factors involving in NO_x formation. (a), Thermal NO_x formation occurs due to the high-temperature

environment in the furnace/boiler [29]. (b), Fuel- NO_x formation comes from the nitrogen-containing species. In this case, ammonia was considered as the fuel [30].

Therefore, NO is accounted for the major portion of NO_x, while NO_2 accounts for below 10% of the total concentration. Thus, the Standard SCR reaction mechanism is the dominant process among the three SCR reaction mechanisms.

According to Kiani et al. (2024), evidence indicates that Standard SCR is best described by the Langmuir-Hinshelwood (L-H) mechanism, which is a process in which both reactants (and/or their activated species) are pre-chemically adsorbed on the catalyst before the reaction occurs. Unlike the L-H mechanism, both the Langmuir-Rideal (L-R) mechanism and the Eley-Rideal (E-R) mechanism are reactions where one chemisorbed species reacts with non-chemisorbed species. The difference between L-R and E-R mechanisms is how such non-chemisorbed species react with chemisorbed species. In the L-R mechanism, the gas-phase species directly collides with the chemisorbed species to initiate the reaction. On the other hand, for the E-R mechanism, the physisorbed species, which are usually formed by van der Waals bonding with the catalyst surface, indirectly react with the chemisorbed species, as shown in the Figure. 7 [31].

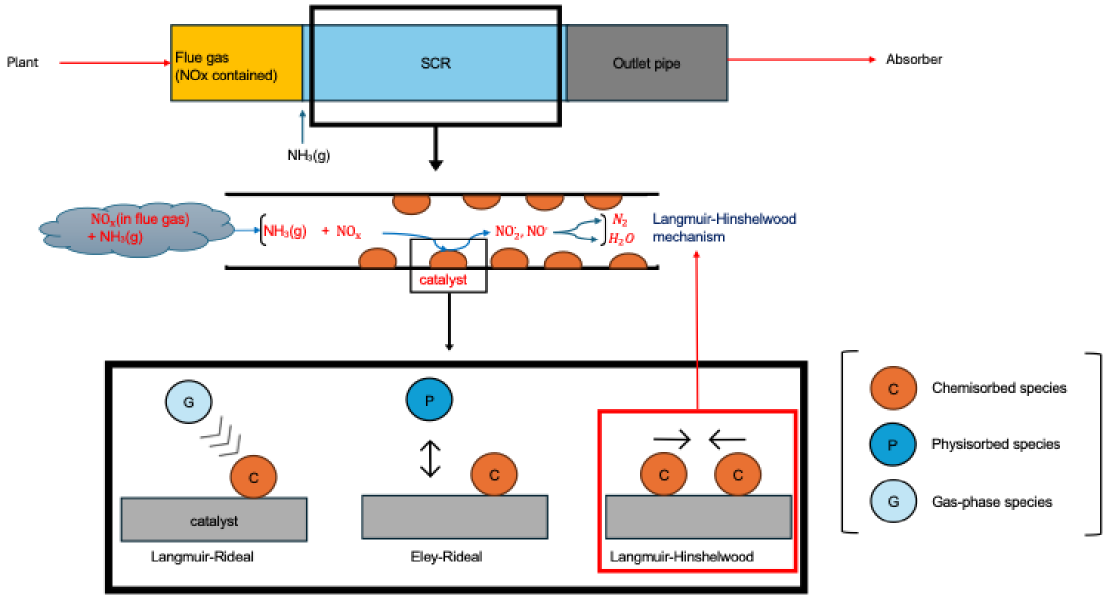


Figure 7. SCR system and its catalytic mechanism. In traditional ammonia-based SCR, ammonia gas is added to the flue gas before it enters the SCR system. Among three different catalytic mechanisms, ammonia SCR is known to follow the Langmuir-Hinshelwood (L-H) mechanism based on spectroscopic evidence and kinetic results. The graphical scheme of those catalysts' mechanisms was extracted from the 2024 work by Kiani and Wachs [31].

In addition, Kiani stated that the L-H mechanism accounts for higher reaction rates due to its secondary dependence on site density, whereas the E-R and L-R mechanisms typically exhibit first-order dependencies and are less commonly dominant pathways [31].

Specifically, the L-H mechanism's second-order dependence on catalytic site density was confirmed by spectroscopic evidence for the presence of surface intermediate species such as NH_2NO. Even though non-idealities, such as site heterogeneity and adsorbate-adsorbate interactions, can seemingly indicate that the kinetic is followed by the E-R mechanism, the L-H mechanism is accepted as the primary route for the ammonia-based SCR. That is the main reason why so many researchers are conducting research to maximize active-site exposure and increase geometric surface area to enhance overall catalytic performance. This could be achieved by considering catalyst structure, such as microstructure, pore texture, and surface area [32].

Regarding catalyst type, the activation temperature of each catalyst is an important consideration for maximizing overall system energy efficiency. According to Table 2, the traditionally used and prevalent catalyst is a Vanadium (V)-based catalyst, which is only useful for power plants that emit high-temperature flue gas [33].

Table 2. Summary of operating conditions for each catalyst. Taken from Wei et al. (2024), ultra-low-temperature SCR catalysts, such as Mn-based and Ce-based catalysts, are suitable for fossil fuel power plants due to their lower operating temperatures compared with conventional V-based, Zeolite, and Fe-based catalysts.

Catalyst Type	Operating Temperature Range (°C)	NO_x Conversion Characteristcs
V-based	Approximately 200- 400	More than 90%
Zeolite catalysts	Approximately 200-600	More than 90%
Fe-based	Approximately 200-580	80-95%
Mn-based	Approximately 100-250	More than 90%
Ce-based	Approximately 150-350	More than 90%

However, such V-based catalysts' operating temperature range doesn't fit well with municipal waste incineration power plants, which usually emit flue gas at around 100-200 ℃. Ultra-low-temperature SCR catalysts, such as Mn-based and Ce-based catalysts, are suitable for use in these types of power plants. Therefore, catalysts should be selected based on the types of plants.

Suggestion of revised Post Carbon Capture system, integrating CESAR-1 with SCR

The distinct advantage of the CESAR-1 is its ability to effectively eliminate CO_2 from flue gas. However, as mentioned in Section 2.2, even a small amount of NO_x in the flue gas can facilitate degradation in the stripper, forming carcinogenic substances, which in turn form additional byproducts that serve as a foundation for further degradation. To minimize this effect, this proposal suggests a revised CESAR-1 system with the SCR method, hereafter referred to as a SCR-CESAR-1 system. (Figure 8.) To briefly introduce our system, flue gas, which typically contains 120,000-150,000 ppm CO_2 and 158 ppm NO_x, is heated to 150℃ - 250 ℃ in the heater. [34], [35] To undergo reaction at low temperature effectively, the catalyst should be chosen among the Ultra-low-temperature SCR catalysts [33]. After NO_x is removed, flue gas is cooled to the absorber's operating temperature (around 40℃) via a heat exchanger and then is introduced into the absorber. The measured temperature of flue gas from coal-fired power plants is around 100-150 ℃. To react NO_x effectively with the catalyst in the SCR Reactor, the heater should be adjusted before the flue gas enters the SCR reactor to deliver a specific amount of energy for reaction. To undergo reaction at low temperatures effectively, the catalyst should be chosen from Ultra-low-temperature SCR catalysts. Once the NO_x is removed from the flue gas, leaving nitrogen gas and water, it should be cooled to the absorber's operating temperature via a heat exchanger, and then the treated flue gas is introduced into the absorber. CO_2 is absorbed with the CESAR-1 solvent to form carbamate, bicarbonate, Protonated-AMP, and Protonated-PZ. CO_2-free gas (clean gas) is emitted to the atmosphere, while CO_2-rich solvent is transported to the stripper for regeneration. During the regeneration process, the CO_2-rich CESAR-1 solvent, in the form of carbamate, bicarbonate, and each protonated amine, is heated with an electrical heater to release CO_2 from the solvent. The detached solvent is captured and returned to the absorber until the solvent's life span is over, and the generated CO_2 is stored in the CO_2 container.

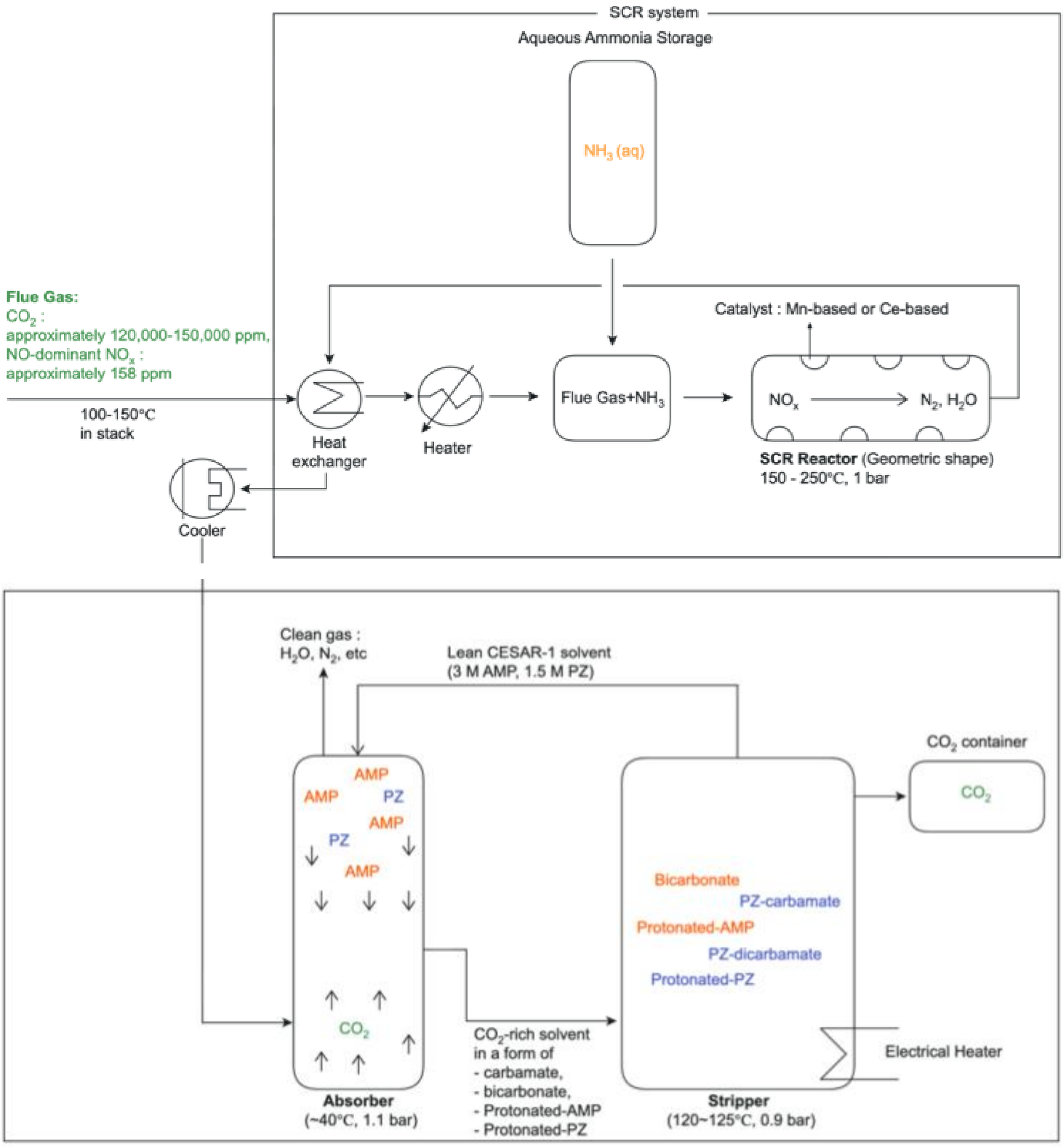


Figure. 8. SCR-CESAR system flow diagram. Flue gas containing CO_2 and NO_x goes into the SCR system. Within the system, a significant amount of NO_x is converted to N_2 and H_2O within the SCR reactor. The remaining gas, which contains non-disrupted CO_2 is cooled and enters the absorber to proceed post-combustion carbon capture process. CO_2 is captured and stored in a container after this reaction.

This combined system (SCR-CESAR system) has several advantages.

First, the SCR–CESAR system is achieved by maintaining similar operating temperatures throughout the entire process—from the thermoelectric power plant to the ammonia-based SCR unit and finally to the CESAR-1 absorber & stripper—while minimizing the effort required for wide temperature changes. This is especially achieved by the usage of ultra-low-temperature SCR catalysts in the design.

Second, the use of Mn-based catalysts not only saves the required energy but also reduces toxicity compared with V-based catalysts, thereby contributing to green chemistry and sustainable industrial processes. [36] Furthermore, the Mn-based catalyst reported that it can be activated in less toxic solvents - such as oxygen molecules and hydrogen peroxide – in which the water is produced as a byproduct, which is also an environmentally friendly substance.

Finally, even though not all the injected ammonia for SCR operation is fully utilized and some remains as ammonia slip, the introduction of an ammonia slip catalyst (ASC) at the end of the SCR system enables the remaining ammonia to react with NO_x. This reduces the need for additional ammonia injections, thereby lowering the overall ammonia consumption and associated costs [37].

Conclusion

Since CO_2 is the greenhouse gas largely produced by human activity, measures to limit these emissions have often been described as an environmental imperative. Among the different types of carbon capture processes, post-combustion capture is widely used, as it can be retrofitted into existing industrial plants and processes. Amine-based absorption using CESAR-1, a blended (AMP) / (PZ) solvent widely applied in post-combustion CO_2 capture, has gained attention due to its higher CO_2 loading capacity and lower regeneration energy, since each amine compensates for the other's weakness.

However, since CO_2 is not the sole constituent of the flue gas, the presence of other constituents, especially NO_x (where x is mostly 1 or 2), in post-combustion CO_2 capture also poses potential problems, including (1) generation of MNPZ, a harmful byproduct to human health and the surrounding aqueous environment, and (2) solvent loss during regeneration. This can be addressed by installing an SCR system upstream of the CESAR-1 process to remove NO_x from the flue stream before it passes through the CESAR-1 absorber. The SCR-CESAR-1 combined system includes the SCR reactor, CO_2 absorber, and stripper that comprises the CESAR-1. The system's advantages include a similar reaction temperature across all stages, which limits heat losses from heating and cooling. The proposed system also uses manganese-based ultra-low-temperature catalysts in the SCR, which are considered less harmful than prevalent yet toxic catalysts such as vanadium.

Additionally, an Ammonia Slip Catalyst (ASC) system can be used to consume unreacted ammonia that may remain after the SCR reaction, thereby reducing the amount of ammonia that must be injected into the flue gas, saving costs and reducing emissions.

For a process or plant with (1) a large NO_x emission, (2) ammonia in the flue gas and (3) temperature of the flue gas within a certain range of 100-200 °C, such as a thermoelectric power plant and a chemical plant, the SCR-CESAR-1 system is particularly suitable. This combined system enables cost-effective, environmentally friendly reaction mechanism for NO_x and CO_2 emission reduction.

Acknowledgements

I would like to express my sincere gratitude to my advisor, Professor Brian M. Stoltz, for his invaluable guidance and insightful feedback. Special thanks to Dong Kyu Chung, a Ph.D. candidate at the California Institute of Technology, for helpful discussions related to this work.

References

[1] Intergovernmental Panel on Climate Change (IPCC). Climate Change: The IPCC 1990 and 1992 Assessments; Overview and Policymaker Summaries and 1992 IPCC Supplement; World Meteorological Organization: Geneva, Switzerland; United Nations Environment Programme: Nairobi, Kenya, 1992; ISBN 0-662-19821-2.
[2] López, J. C.; Quijano, G.; Souza, T. S. O.; Estrada, J. M.; Lebrero, R.; Muñoz, R. Biotechnologies for Greenhouse Gases (CH_4, N_2O, and CO_2) Abatement: State of the Art and Challenges. Appl. Microbiol. Biotechnol. 2013, 97 (6), 2277–2303. https://doi.org/10.1007/s00253-013-4734-z.

[3] Means of Scripps Institution of Oceanography Continuous Data at Mauna Loa and South Pole provided by KenMaarie (personal communication) NASA Goddard Institute for Space Studies (GISS). Global Mean CO_2 Mixing Ratios (ppm): Observations and Future Scenarios. NASA GISS Climate Data;
https://data.giss.nasa.gov/modelforce/ghgases/Fig1A.ext.txt
[4] NOAA Global Monitoring Laboratory. Trends in CO2, NOAA, 2024
https://gml.noaa.gov/webdata/ccgg/trends/co2/co2_mm_mlo.txt.
[5] Schmidt, G. A.; Ruedy, R.; Miller, R. L.; Lacis, A. A. Attribution of the Present-Day Total Greenhouse Effect. J. Geophys. Res. Atmos. 2010, 115(D20), D20106.
https://doi.org/10.1029/2010JD014287
[6] Ros, J.; Agon, N.; Louwagie, D.; Gutiérrez-Sanchez, O.; Huizinga, A.; Gravesteijn, P.; Kruijne, M.; Skylogianni, E.; van Os, P.; Monteiro, J. G. M.CESAR1 Carbon Capture Pilot Campaigns at an Industrial Metal Recycling Site and Analysis of Solvent Degradation Behavior.Ind. Eng. Chem. Res. 2025, 64(10), 5548–5565.https://doi.org/10.1021/acs.iecr.4c03998
[7] Lv, B.; Guo, B.; Zhou, Z.; Jing, G. Mechanisms of CO_2 Capture into Monoethanolamine Solution with Different CO_2 Loading during the Absorption/Desorption Processes. Environ. Sci. Technol. 2015, 49 (17), 10728–10735.
https://doi.org/10.1021/acs.est.5b02356
[8] Zhang, Z.; Vo, D.-N.; Kum, J.; Hong, S.-H.; Lee, C.-H. Enhancing Energy Efficiency of Chemical Absorption-Based CO_2 Capture Process with Advanced Waste-Heat Recovery Modules at a High Capture Rate. Chem. Eng. J. 2023, 472, 144918.
https://doi.org/10.1016/j.cej.2023.144918.
[9] Wang, N.; Wang, D.; Krook-Riekkola, A.; Ji, X. MEA-Based CO_2 Capture: A Study Focuses on MEA Concentrations and Process Parameters. Front. Energy Res. 2023, 11, 1230743. https://doi.org/10.3389/fenrg.2023.1230743.
[10] Hasan, S.; Abbas, A. J.; Nasr, G. G. Improving the Carbon Capture Efficiency for Gas Power Plants through Amine-Based Absorbents. Sustainability 2021, 13 (1), 72.
https://doi.org/10.3390/su13010072
[11] Bottoms, R. R. Process for Separating Acidic Gases. U.S. Patent 1,783,901, Dec 2, 1930.
[12] Vinjarapu, S. H. B.; Neerup, R.; Larsen, A. H.; Jørsboe, J. K.; Villadsen, S. N. B.; Jensen, S.; Karlsson, J. L.; Kappel, J.; Lassen, H.; Blinksbjerg, P.; von Solms, N.; Fosbøl, P. L.
Results from Pilot-Scale CO_2 Capture Testing Using 30 wt % MEA at a Waste-to-Energy Facility: Optimisation through Parametric Analysis. Appl. Energy 2024, 355, 122193. https://doi.org/10.1016/j.apenergy.2023.122193.
[13] Zhang, W.; Liu, H.; Sun, C.; Cakstins, J.; Sun, C.; Snape, C. E. Parametric Study on the Regeneration Heat Requirement of an Amine-Based Solid Adsorbent Process for Post-Combustion Carbon Capture. Appl. Energy 2016, 168, 394–405.
https://doi.org/10.1016/j.apenergy.2016.01.049.
[14] Haimour, N.; Sandall, O. C. Absorption of Carbon Dioxide into Aqueous Methyldiethanolamine. Chem. Eng. Sci. 1984, 39 (12), 1791–1796.
https://doi.org/10.1016/0009-2509(84)80115-3.
[15] Mackenzie, D. H.; Prambil, F. C.; Daniels, C. A.; Bullin, J. A. Design and Operation of a Selective Sweetening Plant Using MDEA. Energy Progress 1987, 7 (1), 31–36.
[16] Mertdag, Y. Process Simulation and Techno-Economic Analysis of CESAR-1 Solvent in Post Combustion Carbon Capture, and Benchmarking with MEA Solvent; Master Thesis, Aalborg University, 2023.
[17] Smahi, A.; Kanniche, M.; Bouallou, C.; Jourdan, N.; Authier, O. CO_2 Capture in MDEA–PZ and AMP–PZ Solvents: ELECNRTL vs ENRTL-RK Models and Their Efficiency in Natural Gas Combined Cycles. Proc. 17th Int. Conf. on Greenhouse Gas Control Technologies (GHGT-17), Calgary, Canada, Oct 20–24, 2024. Available at SSRN: https://ssrn.com/abstract=5025846
[18] Ros, J.; Agon, N.; Louwagie, D.; Gutiérrez-Sanchez, O.; Huizinga, A.; Gravesteijn, P.; Kruijne, M.; Skylogianni, E.; van Os, P.; Moretz-Sohn Monteiro, J. G. M. CESAR1

Carbon Capture Pilot Campaigns at an Industrial Metal Recycling Site and Analysis of Solvent Degradation Behavior. Ind. Eng. Chem. Res. 2025, 64 (10), 5548–5565. https://doi.org/10.1021/acs.iecr.4c03998
[19] Morgan, J. C.; Campbell, M.; Putta, K. R.; Shah, M. I.; Matuszewski, M.; Omell, B. Development of Process Model of CESAR1 Solvent System and Validation with Large Pilot Data. Presented at the 16th International Conference on Greenhouse Gas Control Technologies (GHGT-16), Lyon, France, October 23–27, 2022.
[20] Stowe, H. M.; Vilčiauskas, L.; Paek, E.; Hwang, G. S. On the Origin of Preferred Bicarbonate Production from Carbon Dioxide (CO_2) Capture in Aqueous 2-Amino-2-Methyl-1-Propanol (AMP). Phys. Chem. Chem. Phys. 2015, 17, 29184–29192. https://doi.org/10.1039/C5CP04876A.
[21] Conway, W.; Fernandes, D.; Beyad, Y.; Burns, R.; Lawrance, G.; Puxty, G.; Maeder, M. Reactions of CO_2 with Aqueous Piperazine Solutions: Formation and Decomposition of Mono- and Dicarbamic Acids/Carbamates of Piperazine at 25.0 °C. J. Phys. Chem. A 2013, 117, 806–813. https://doi.org/10.1021/jp310560b.
[22] Li, T.; Yu, Q.; Barzagli, F.; Li, C.; Che, M.; Zhang, Z.; Zhang, R. Energy Efficient Catalytic CO_2 Desorption: Mechanism, Technological Progress and Perspective. Carbon Capture Sci. Technol. 2023, 6, 100099. https://doi.org/10.1016/j.ccst.2023.100099
[23] Ros, J.; Agon, N.; Louwagie, D.; Gutiérrez-Sanchez, O.; Huizinga, A.; Gravesteijn, P.; Kruijne, M.; Skylogianni, E.; van Os, P.; Moretz-Sohn Monteiro, J. G. CESAR1 Carbon Capture Pilot Campaigns at an Industrial Metal Recycling Site and Analysis of Solvent Degradation Behavior. Ind. Eng. Chem. Res. 2025, 64, 5548-5565. https://doi.org/10.1021/acs.iecr.4c03998.
[24] Languille, B.; Drageset, A.; Mikoviny, T.; Zardin, E.; Benquet, C.; Ullestad, Ø.; Aronson, M.; Kleppe, E. R.; Wisthaler, A. Atmospheric emissions of amino-methyl-propanol, piperazine and their degradation products during the 2019-20 ALIGN-CCUS campaign at the Technology Centre Mongstad. Presented at the 15th International Conference on Greenhouse Gas Control Technologies (GHGT-15), Abu Dhabi, UAE, March 15–18, 2021.
[25] Buvik, V.; Vernstad, K.; Grimstvedt, A.; Høisæter, K. K.; Vevelstad, S. J.; Knuutila, H. K. Degradation of 2-Amino-2-Methylpropanol and Piperazine at CO_2 Capture-Relevant Conditions. Ind. Eng. Chem. Res. 2025, 64 (22), 11000–11020. https://doi.org/10.1021/acs.iecr.5c00527
[26] Vikram, H. P. R.; Kumar, T. P.; Kumar, G.; Deka, R.; Suhail, S. M.; Jat, S.; Bannimath, N.; Padmanabhan, G.; Chandan, R. S.; Kumar, P.; Gurupadayya, B. Nitrosamines Crisis in Pharmaceuticals—Insights on Toxicological Implications, Root Causes and Risk Assessment: A Systematic Review. J. Pharm. Anal. 2024, 14 (5), 100919. https://doi.org/10.1016/j.jpha.2023.12.009.
[27] Iwasaki, M.; Shinjoh, H. A Comparative Study of “Standard”, “Fast” and “NO_2” SCR Reactions over Fe/Zeolite Catalyst. Appl. Catal. A: Gen. 2010, 390, 71–77. https://doi.org/10.1016/j.apcata.2010.09.034.
[28] Saito, T.; Fujiwara, K. Causal analysis of nitrogen oxides emissions process in coal-fired power plant with LiNGAM. Front. Anal. Sci. 2023, 3, 1045324. https://doi.org/10.3389/frans.2023.1045324
[29] Miller, J. A.; Bowman, C. T. Mechanisms of Nitrogen Oxide Formation in Combustion Processes. Combust. Flame 1989, 76, 1–31. https://doi.org/10.1016/0360-1285(89)90017-8
[30] Zevenhoven, R.; Kilpinen, P. NITROGEN — Mechanisms and Kinetics of Nitrogen Compounds in Combustion; Åbo Akademi University: Finland, 2002; Chapter 4.
[31] Kiani, D.; Wachs, I. E. Practical Considerations for Understanding Surface Reaction Mechanisms Involved in Heterogeneous Catalysis. ACS Catal. 2024, 14 (22), 16770–16784. https://doi.org/10.1021/acscatal.4c05188
[32] Trovarelli, A.; Llorca, J. Catalysis Today 1998, 39 (2), 141-153. https://doi.org/10.1016/S0920-5861(98)00050-9.
[33] Wei, F.; Rao, Y.; Huang, Y.; Wang, W.; Mei, H.; et al. The New Challenges for the

Development of NH3-SCR Catalysts under New Situation of Energy Transition in Power Generation Industry. Chin. Chem. Lett. 2024, 35, 108931. https://doi.org/10.1016/j.cclet.2023.108931.
[34] Songolzadeh, M.; Soleimani, M.; Takht Ravanchi, M.; Songolzadeh, R. Carbon Dioxide Separation from Flue Gases: A Technological Review Emphasizing Reduction in Greenhouse Gas Emissions. The Scientific World Journal 2014, 2014, 828131. https://doi.org/10.1155/2014/828131
[35] Li, Z.; Fan, T. W.; Lun, M. S.; Li, Q. Optimization of Municipal Solid Waste Incineration for Low-NOx Emissions through Numerical Simulation. Sci. Rep. 2024, 14, 19309. https://doi.org/10.1038/s41598-024-69019-w.
[36] Kazemi, M.; Ali, R.; Jain, V.; Ballal, S.; Abosaoda, M. K.; Singh, A.; Krithiga, T.; Joshi, K. K.; Javahershenas, R. A Comprehensive Review on Magnetic Manganese as Catalysts in Organic Synthesis. RSC Adv. 2025, 15, 23054. https://doi.org/10.1039/d5ra02493e.
[37] Manassero, D.; Bonvicini, F.; Ravelli, D.; Malavasi, L.; Carnelli, L.; Arlotti, G.; Pellacani, G. Pure Ammonia-Fueled Roller Kiln for the Production of Ceramic Tiles: A First Demonstration. Energy Fuels 2024, 38 (7), 6355–6362. https://doi.org/10.1021/acs.energyfuels.4c03745.

Authors

Dongseok Kim is an undergraduate student at Queens College, City University of New York, majoring in Chemistry. His academic interests include heterogeneous catalysis, particularly ultra-low-temperature catalysts for industrial applications.

The Impact of Query Decomposition and Cross-Encoder Reranking in Multi-Hop Retrieval-Augmented Generation

Hail Lim, Michigan State University, 132 Crossroads Ln., Troy, Michigan, 48083, United States
Corresponding author email: ianlim.dev@gmail.com

Abstract

Retrieval-Augmented Generation (RAG) has emerged as a promising paradigm for open-domain question answering. However, standard single-hop retrieval often fails on complex, multi-hop queries where the answer requires synthesizing information from disparate documents. In this work, we propose an enhanced Multi-Hop RAG pipeline augmented with Cross-Encoder Reranking to address the challenges of reasoning across multiple documents. Our approach decomposes complex queries into self-contained sub-questions and employs a Cross-Encoder to rerank candidates at each retrieval step, mitigating the "semantic drift" inherent in dense vector search. We systematically evaluate our system against two baselines—Standard Single-Hop RAG and Decomposed Multi-Hop RAG—using a curated subset of the HotpotQA dataset. Experimental results demonstrate that our proposed method achieves superior accuracy (62%, a 20% gain over the single-hop baseline) by effectively filtering distractors. Furthermore, our ablation studies reveal a fundamental "Recall Ceiling" in dense retrieval, where blindly increasing the candidate pool yields diminishing returns. Based on these findings, we identify a "Wide Net, Tight Filter" strategy as the Pareto-optimal configuration for balancing reasoning accuracy with system latency.

Keywords

Large language models (LLM); Retrieval-augmented generation (RAG); Multi-hop question; Query decomposition; Reranking; HotpotQA.

Introduction

Retrieval-Augmented Generation (RAG) [1,2,3] has become the de facto standard for equipping Large Language Models (LLMs) with up-to-date, non-parametric knowledge. While standard RAG pipelines excel at "Single-Hop" queries—where the answer resides in a semantically similar passage—they often falter in realistic, complex scenarios. Many information-seeking tasks are inherently Multi-Hop [4,5,6], requiring the synthesis of facts dispersed across disparate documents connected by implicit reasoning chains rather than explicit keyword overlap [7,8,9,10]. In such cases, the semantic vector of the original question aligns poorly with the document containing the final answer, leading to "semantic drift".

Despite the proliferation of RAG frameworks, architectural best practices for these complex queries remain under-documented. Practitioners face a dilemma: Standard dense retrieval (Bi-Encoder) is efficient but lacks the reasoning depth to bridge semantic gaps; conversely, complex "Agentic" workflows involving iterative decomposition or heavy cross-encoder reranking promise higher accuracy but introduce significant latency [11,12,13,14]. Furthermore, simply retrieving more documents often yields diminishing returns due to the "Recall Ceiling" of dense embeddings—a fundamental limitation frequently overlooked in favor of generative optimizations [15,16,17,18].

Therefore, bridging the gap between naive RAG and robust multi-hop reasoning requires a rigorous empirical understanding of design trade-offs. Specifically: *When does the cost of query decomposition pay off? Is the precision gain from a cross-encoder worth the latency overhead? And at what point does retrieval recall become the insurmountable bottleneck?*

In this work, we study two such design decisions that are particularly salient in practice:

1. **Should we explicitly decompose complex questions?** Prompting LLMs to decompose questions into simpler sub-questions is theoretically sound but introduces sequential dependencies. It is not obvious under which conditions this complexity outperforms direct retrieval.
2. **Should we add a cross-encoder reranking stage?** Cross-encoders capture subtle logical relations by jointly encoding query-document pairs [], unlike interaction-light Bi-Encoders. For multi-hop RAG, this raises a critical trade-off between answer quality and system latency, especially when reranking is performed iteratively.

Our goal is to make these trade-offs concrete and measurable in a realistic, controlled setting. We build a modular RAG pipeline on top of **HotpotQA** [4] and systematically compare three baselines:

Baseline A: Single-hop RAG. Retrieves top-k passages for the original question.
Baseline B: Multi-hop RAG with Decomposition. Decomposes the question and retrieves evidence sequentially.
Baseline C: Multi-hop RAG with Cross-Encoder Reranking. Adds a reranking stage to filter distractors from a larger candidate pool.

Concretely, we investigate the efficacy and limitations of these architectures by addressing three interconnected dimensions. First, we examine Decomposition Effectiveness, quantifying how well decomposition mitigates semantic drift. Second, we analyze the Cost-Benefit Ratio of Reranking, determining if precision gains justify the computational latency. Finally, we explore System Boundaries, specifically characterizing the "Recall Ceiling"—the saturation point where increasing retrieval hyperparameters no longer yields accuracy improvements.

In summary, the contribution of this paper are three-fold:

1. We design and open-source a modular Multi-Hop RAG pipeline that cleanly separates indexing, decomposition, retrieval, and reranking.
2. We perform a systematic evaluation, revealing a performance hierarchy where our Reranking proposal achieves a 20% accuracy gain over the single-hop baseline.
3. We provide actionable engineering guidelines, identifying a "Wide Net, Tight Filter" strategy (large number pre-fetch candidates, small number of retrievals) as the Pareto-optimal configuration, while characterizing the fundamental "Recall Ceiling" on hard queries.

Background & Related Work

Large language models (LLMs) are rapidly transforming the landscape of artificial intelligence. With their expanding capabilities in language understanding and generation, LLMs have been widely adopted in knowledge-intensive scenarios, including everyday question answering, engineering workflows, and scientific research assistants. However, even very large models have a fixed training cutoff and may lack access to up-to-date or domain-specific information, and it is often impractical or impossible to continually retrain them on all relevant data sources. This has created a growing demand for techniques that integrate LLMs with external knowledge in a scalable and controllable way.

1. Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) has emerged as a standard solution to this challenge [21, 22, 23, 24, 25]. In RAG, a parametric language model is coupled with a non-parametric memory, typically a dense index over a text corpus, and generation is conditioned on documents retrieved from that index. Lewis et al [1]. introduced RAG as a general-purpose recipe for knowledge-intensive NLP tasks, showing that augmenting a seq2seq model with a learned neural retriever over Wikipedia improves performance and reduces hallucination compared to purely parametric models. Subsequent work and system reports have documented wide adoption of RAG in industrial question answering, document assistants, and domain-specific copilots. Conceptually, many early RAG systems follow a two-stage workflow very similar to the one we adopt in this paper:

1. A retrieval stage, which gathers context-relevant passages from an external knowledge base;
2. A generation stage, which conditions the LLM on the retrieved passages (e.g., by concatenating them into the prompt) to produce answers that are more accurate and grounded.

This architecture enables models to access information beyond their original training data without incurring the prohibitive cost of full retraining and allows practitioners to update knowledge by re-indexing documents rather than modifying model weights. Empirically, retrieval-augmented systems have been shown to reduce hallucinations by grounding responses in explicit evidence and to support stricter data-governance regimes, since sensitive documents can be controlled at the retrieval layer rather than embedded directly in model parameters.

2. Multi-Hop Question Answering and HotpotQA

While RAG has become the default for single-hop or factoid QA, multi-hop questions pose additional challenges. In multi-hop QA [4,5,6,27,28], a system must gather and integrate information from multiple documents to answer a question; solving the task requires both accurate retrieval and explicit or implicit reasoning across pieces of evidence.

HotpotQA [4] is a widely used benchmark for this setting. Yang et al. [4] introduced HotpotQA as a dataset of 113k Wikipedia-based questions requiring reasoning over multiple supporting documents, along with sentence-level supporting facts for strong supervision and explainable predictions. The dataset includes “comparison” questions and distractor paragraphs, making it a realistic and challenging environment for multi-hop retrieval and reasoning. HotpotQA has spurred a line of work on pipeline architectures that combine document selection, sentence selection, and reading comprehension models, as well as graph-based and neural-reasoning approaches to multi-document QA. In the era of LLMs and RAG, HotpotQA remains a natural testbed for evaluating whether retrieval-augmented LLMs can perform explicit multi-hop reasoning when provided with a large but noisy evidence pool.

3. LLM-Based Decomposition and Multi-Step Reasoning

The emergence of instruction-tuned LLMs has enabled new approaches to multi-step reasoning that rely heavily on prompting [7,8,9]. Chain-of-Thought (CoT) [10] prompting demonstrates that providing or eliciting intermediate reasoning steps can significantly improve performance on arithmetic, common sense, and symbolic reasoning tasks [23]. More broadly, recent work on tool-using agents and planner–executor architectures shows that LLMs can be prompted to decompose complex tasks into sub-tasks and call external tools such as search APIs in a loop, effectively implementing a form of decompose–solve–compose reasoning.

In the context of multi-hop QA and RAG, these ideas have inspired methods that decompose a complex question into a sequence of sub-questions, retrieve and answer each sub-question separately, and then synthesize a final answer from the intermediate

results. Such decomposition is attractive because it (i) makes retrieval queries more focused, potentially improving recall, and (ii) yields interpretable intermediate steps that can be inspected or debugged. However, decomposition also introduces new failure modes: if the decomposition is incorrect or misaligned with the underlying reasoning structure, downstream retrieval and answering may be led astray.
Our multi-hop baselines are directly motivated by this line of work. We use an LLM to generate a sequence of sub-questions from the original query, then apply RAG per sub-question, and finally synthesize a final answer. This allows us to empirically study when decomposition improves performance over a single-hop RAG baseline on HotpotQA-style questions.

Methods

In this study, we developed and evaluated three distinct RAG pipelines to investigate the efficacy of multi-hop reasoning and cross-encoder reranking in answering complex, bridge-type questions. All pipelines were implemented using the LangChain framework, utilizing OpenAI's gpt-3.5-turbo as the generator and text-embedding-ada-002 for dense retrieval. The document corpus consists of paragraphs from the HotpotQA dataset, indexed using a FAISS vector store [26].

1. Baseline 1: Single-Hop RAG (Standard Baseline)

The Single-Hop RAG serves as the naive baseline, representing standard industry practices where retrieval is performed once based on the original user query. Given a complex user query Q, this pipeline performs a direct dense retrieval against the vector index V. The system computes the cosine similarity between the query embedding $E(Q)$ and document embeddings $E(d_i$) for all documents in the corpus. The top-k documents

$$\mathcal{D}_{retrieved} = \text{Top-}k\left(\cos(E(Q), E(d_i))\right)$$

with the highest similarity scores are retrieved:
In our experiments, we set $k = 5$. These retrieved documents are concatenated into a context string C and passed to the LLM along with the original query to generate the final answer A.
Limitation. This baseline assumes that the answer to a complex question resides within a single retrieval step's semantic vicinity. However, for multi-hop questions (e.g., "Who is the director of the film starring X?"), the necessary information for the second hop ("film starring X") may not share high semantic overlap with the original query, often leading to retrieval failure.

2. Baseline 2: Multi-Hop RAG via Query Decomposition

To address the limitations of single-hop retrieval, we implemented a Multi-Hop RAG pipeline that mimics the human reasoning process by decomposing complex questions into a sequence of simpler, self-contained sub-questions. We explain the workflow below.

Query Decomposition Mechanism. We employ a "Decompose-then-Execute" strategy. The core component is a decomposition chain driven by a specifically engineered prompt. The prompt instructs the LLM to break down the input question Q into a list of sub-questions $\{q_1, \ldots, q_n\}$. Crucially, to prevent information loss during independent retrieval steps, we enforce a Context Retention Constraint in our decomposition prompt. This ensures that sub-questions are self-contained and do not rely on vague pronouns (e.g., resolving "When was he born?" to "When was the creator of The Simpsons born?"). The decomposition prompt template is defined as follows:
You are an expert at query decomposition for retrieval systems. Break down the complex question into a series of simple sub-questions. IMPORTANT: Each sub-question must be SELF-CONTAINED. Do NOT use vague pronouns without referring to the original context.

Iterative Retrieval and Synthesis. The pipeline executes sequentially. *(1) Decomposition:* The LLM generates the list of sub-questions. *(2) Sequential Retrieval:* For each sub-question q_j, the system performs an independent dense retrieval to obtain a specific context set D_j. *(3) Synthesis:* All retrieved contexts from all steps are aggregated. A final synthesis prompt guides the LLM to reason across these disjoint pieces of information to construct the final answer.

3. Proposal (Baseline 3): Multi-Hop RAG with Cross-Encoder Reranking

While Baseline 2 improves reasoning, dense retrieval (Bi-Encoder) often suffers from "semantic drift," retrieving documents that are semantically similar but factually irrelevant (distractors). To mitigate this, we propose our primary method: a Multi-Hop pipeline augmented with a Cross-Encoder Reranker.

Reranking Mathematical Formulation. In the standard retrieval stage (Bi-Encoder), the relevance score s_{bi} is the dot product of independent vectors: $s_{bi}(q,d) = \varphi(q)^T \psi(d)$, where and ψ are separate encoding networks (or the same network). This is computationally efficient but lacks deep interaction between query and document tokens. In our proposal, we introduce a re-ranking stage using a Cross-Encoder model M_{cross}, specifically ms-marco-MiniLM-L-6-v2. Unlike the Bi-Encoder, the Cross-Encoder takes the concatenated pair of the query and document as a single input, allowing full self-

$$s_{cross}(q,d) = \text{Sigmoid}\left(\mathcal{M}_{cross}([CLS] \circ q \circ [SEP] \circ d)\right)$$

attention across both sequences:
where *CLS* is the classification token output used to predict the relevance probability.

Pipeline Implementation. The enhanced retrieval process for each sub-question q_j operates as follows. *(1) Recall Expansion:* We first retrieve a larger set of candidates C (where $|C| = K_{candidates} = 15$) using the standard vector search. *(2) Scoring:* Each candidate document $d \in C$ is paired with the sub-question q_j and scored using the Cross-Encoder to obtain $s_{cross}(q_j, d)$. *(3) Reranking & Pruning:* Documents are sorted by s_{cross} in descending order. We retain the top-K' documents (where $K' = 7$) to form the final refined context.

$$\mathcal{D}_{final} = \{d \in \mathcal{C} \mid \text{Rank}(s_{cross}(q_j, d)) \leq K'\}$$

This reranking step acts as a semantic filter, effectively removing high-similarity distractors (e.g., documents sharing keywords but different entities) and ensuring that the subsequent generation step is grounded in high-precision evidence.

Experiment Setup

To rigorously evaluate the proposed Multi-Hop RAG with Reranking against standard baselines, we designed a comprehensive experimental workflow encompassing knowledge base construction, automated test set curation, and LLM-based evaluation. The implementation details are available in our open-source repository, specifically within *config.py* and *rag_engine.py*.

1. Dataset and Knowledge Base Construction. We utilized the **HotpotQA** dataset (distractor setting), a benchmark specifically designed for multi-hop reasoning. To simulate a realistic, resource-constrained RAG environment, we constructed a local knowledge base using a dense subset of the training split (e.g. train[:1000]). The document processing pipeline (*data_loader.py* and *vector_store.py*) involved the following steps:

(1) *Chunking:* Raw documents were split into chunks of 256 tokens with a 100-token overlap to preserve contextual continuity across boundaries.

(2) *Indexing:* We generated dense vector embeddings using OpenAI's `text-embedding-ada-002`. These embeddings were indexed using FAISS (Facebook AI Similarity Search) to enable efficient similarity retrieval.

This process will output files *index.faiss* and *index.pkl* for future reference so that we do not need to generate them in each experiment.

2. Automated Test Set Curation. To ensure the evaluation focused on reasoning capabilities rather than simple keyword matching, we did not random sample questions. Instead, we implemented a targeted curation script (*generate_test_set.py*). We filtered the dataset to select questions meeting two specific criteria: *(1) Type Constraint* (type="bridge"). This ensures the question requires connecting multiple distinct pieces of information (e.g., finding a director, then finding their birth year). (2) *Difficulty Constraint:* (level="medium"). We selected medium-difficulty questions to balance the need for multi-hop reasoning while avoiding obscure entities that might suffer from embedding recall failures in a small index. This process yielded a curated test set where the ground truth answers are guaranteed to be present in the indexed corpus, isolating the retrieval and reasoning performance of the models.

3. Experimental Baselines. We compared three distinct RAG configurations using the same underlying Generator (`gpt-3.5-turbo`) and Vector Store. The construction of these baselines are elaborated in the Methods section. Here we focus on the setup parameters.
Baseline 1: Single-Hop RAG. A naive approach performing a single vector search ($K = 8$) followed by answer generation. This serves as the lower bound for performance.
Baseline 2: Multi-Hop RAG. A decompose-then-execute pipeline where the complex query is broken into sub-questions. Each sub-question triggers a standard vector search.
Baseline 3 (Proposal): Multi-Hop RAG with Reranking. It augments the multi-hop pipeline with a "Recall-then-Precision" strategy: *(1) Recall.* It fetches a larger pool of candidates ($K_{candidate} = 30$). *(2) Rerank:* A Cross-Encoder (`ms-marco-MiniLM-L-12-v2`) scores the relevance of each candidate. *(3) Safety Fusion:* To prevent over-pruning, we force the inclusion of the top-1 vector search result alongside the highest-scoring reranked documents before generation.

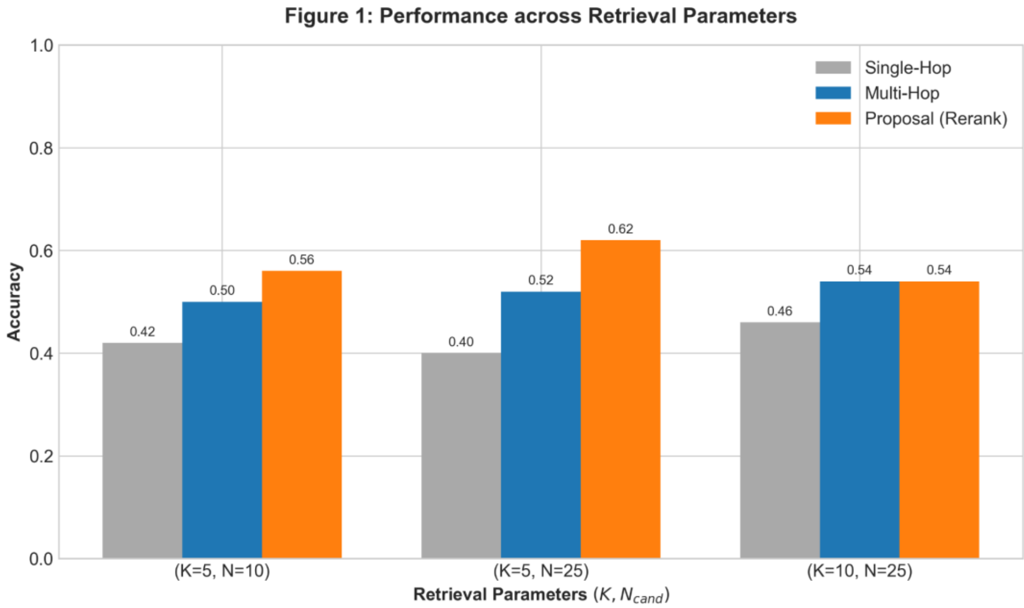
4. Evaluation Protocol: LLM-as-a-Judge
Given the generative nature of the answers, exact string matching is insufficient for accuracy calculation. We implemented an automated evaluation pipeline (*evaluator.py*) utilizing **LLM-as-a-Judge** [19,20]. To ensure reliability and mitigate the "lazy judge" phenomenon often seen in smaller models, we employed GPT-4o as the evaluator. The judge operates with a Chain-of-Thought prompt that enforces a strict verification protocol that extracts key entities (names, dates, locations) from the Ground Truth, verifies their presence in the Generated Answer, and penalizes vague answers or hallucinations.
The system assigns a binary score (Correct/Incorrect) to each output. Final accuracy is reported as the percentage of correct answers across the test set. All experimental parameters, including model names and retrieval hyperparameters, are logged automatically to *config.py* to ensure reproducibility.

Results & Discussion

In this section, we provide a comprehensive analysis of the experimental results. We evaluate the proposed Multi-Hop RAG with Reranking against the baselines across three critical dimensions: sensitivity to retrieval hyperparameters, computational efficiency trade-offs, and robustness across varying difficulty levels.

Performance under Varying Retrieval Constraints

We first investigate how the three RAG pipelines behave under different retrieval configurations. Specifically, we manipulate the Recall depth (N_{cand}, the number of candidates initially fetched) and the Precision scope (K, the final context window size passed to the LLM). As illustrated in **Figure 1**, a consistent performance hierarchy emerges in the optimal setting ($\boldsymbol{N_{cand} = 25, K = 5}$): the Proposal (Multi-Hop + Reranking) significantly outperforms the Multi-Hop baseline, which in turn surpasses the Single-Hop baseline. This confirms that complex bridge questions suffer from a semantic gap that cannot be bridged by a single vector search step.



A more nuanced phenomenon becomes apparent when we perform an ablation study on the hyperparameter constraints.

a. The "Context Saturation" Effect ($\boldsymbol{N_{cand} = 25, K = 10}$). When we relax the final context constraint by increasing K to 10, we observe that the performance gap between the Proposal and the Multi-Hop baseline narrows. This convergence suggests that GPT-3.5 possesses a latent capacity to perform "internal reranking"; if the vector search manages to place the correct document anywhere within the top 10 results, the LLM can often locate it without explicit external reranking. This indicates that the architectural value of the Cross-Encoder is most pronounced in resource-constrained environments (small K), where it acts as a critical filter to maximize the information density of the context window.

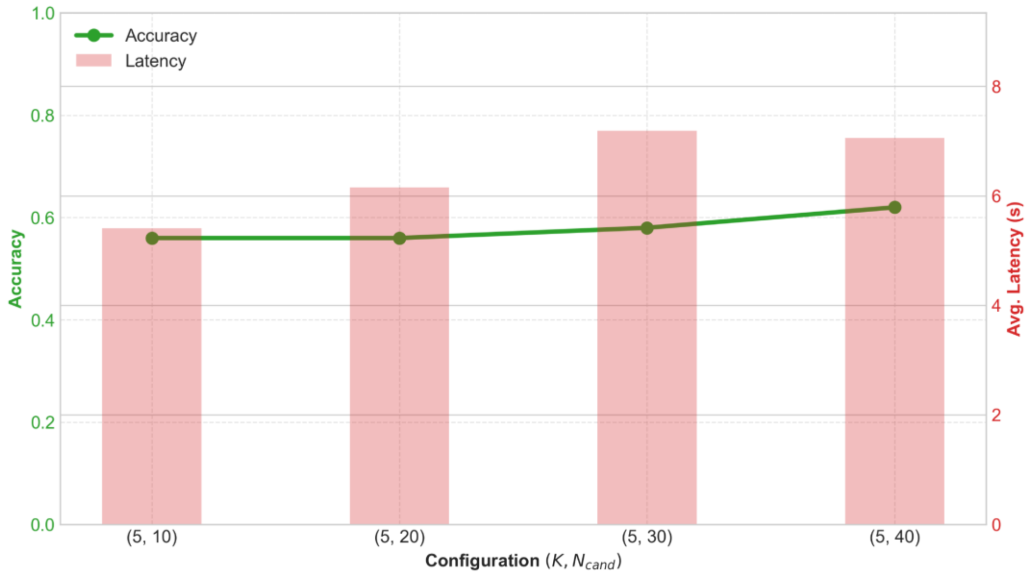
b. The "Recall Bottleneck" Effect ($\boldsymbol{N_{cand} = 10, K = 5}$). Conversely, when we restrict the initial fetch size to 10, the Proposal's performance degrades, failing to significantly outperform the baseline. This reveals a "Recall Bottleneck." Due to semantic drift, the correct document often falls outside the top-10 boundary in the vector space. If the initial fetch size is too small, the relevant document is pruned before the Reranker even has a chance to score it. This empirically validates our design choice of a "Wide Net, Tight Filter" strategy ($N >> K$), ensuring that the Reranker has a sufficient candidate pool to exercise its discriminative power.

The Efficiency Trade-off: Diminishing Returns of Recall

A key engineering challenge in deploying RAG systems is balancing recall quality with system latency. **Figure 2** visualizes this trade-off for the Proposal model, plotting accuracy against average query latency as the initial fetch size (N_{cand}) increases from 10 to 40, while keeping the retrieval window (K) fixed.

Contrary to the expectation that a wider search radius would yield proportional performance gains, the data reveals a distinct asymptotic trend in accuracy that contrasts sharply with the linear increase in latency. As depicted in the figure, inference time grows linearly with N_{cand} due to the computational overhead of the Cross-Encoder scoring every additional candidate pair. However, the accuracy curve exhibits an early saturation pattern. Increasing the fetch size from 10 to 20 yields modest improvements or even worse performance, and extending beyond 30 results in statistically negligible marginal gains. This behavior indicates a fundamental "Recall Ceiling" inherent to the dense embedding model. For this specific domain, the retrieval outcome tends to be binary: either the relevant document is semantically close enough to appear in the top tier of results, or it suffers from severe semantic drift and falls outside the retrieval horizon entirely. There is little evidence of a "middle ground" where correct documents are consistently buried between rank 30 and 40. Consequently, blindly increasing the prefetch buffer beyond N_{cand} introduces significant computational waste without tangible

Figure 2: Accuracy vs. Latency Trade-off (Proposal)



benefits. Based on this cost-benefit analysis, we identify $N_{cand} = 20$ as the Pareto-optimal configuration, balancing robust recall with acceptable system latency.

Robustness and Failure Mode Analysis

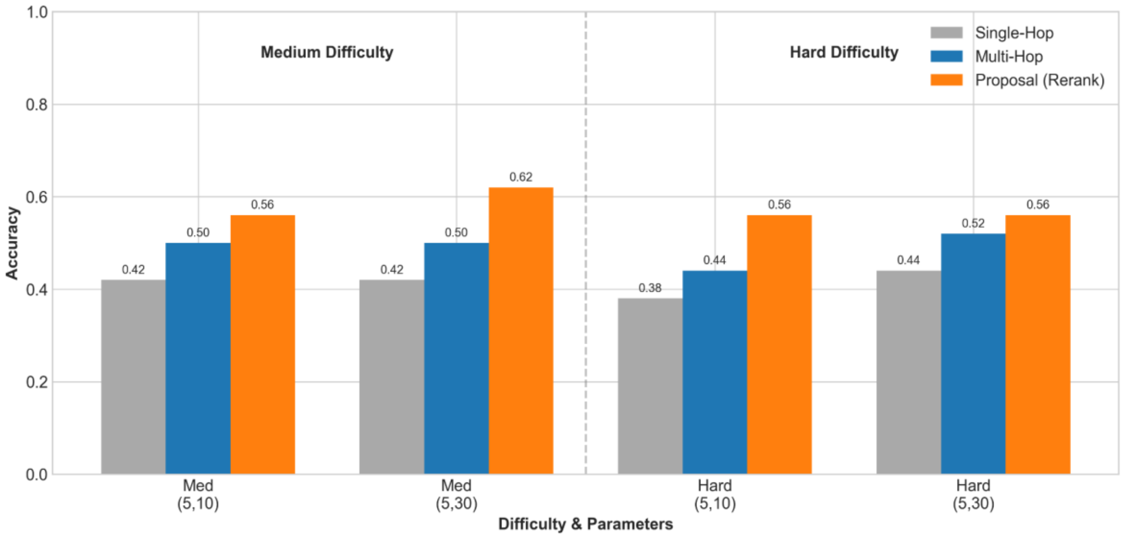
Finally, we assessed the generalization capabilities of the models by stratifying the test set into "Medium" and "Hard" difficulty levels. **Figure 3** depicts the performance degradation observed across all models when transitioning to harder questions.

It is notable that while our Proposal maintains its lead in the Medium category, the performance advantage over the Multi-Hop baseline diminishes in the Hard category. This result points to a critical "Garbage In, Garbage Out" bottleneck in the RAG pipeline. Hard questions in HotpotQA often involve obscure entities or highly indirect relationships that challenge the initial query decomposition step. If the first-hop retrieval fails to recall the necessary "bridge" document—a scenario significantly more probable with Hard questions—the Reranker receives a candidate list consisting entirely of noise. Since a

Reranker is a discriminative model rather than a generative one, it cannot recover information that was never retrieved. This finding highlights that while Cross-Encoder Reranking is a powerful tool for filtering distractors and improving precision, it cannot compensate for fundamental failures in the initial Recall stage. Future optimizations for hard queries should thus prioritize upstream improvements, such as hybrid search (BM25 [] + Dense) or enhanced query reformulation strategies, rather than simply scaling the downstream reranking component.

Conclusion & Outlook

Figure 3: Robustness Analysis (Medium vs. Hard Difficulty)



In this work, we presented a robust Multi-Hop RAG pipeline augmented with Cross-Encoder Reranking to tackle complex, bridge-type questions. By systematically evaluating our approach against standard baselines on the HotpotQA dataset, we demonstrated that query decomposition is a prerequisite for multi-step reasoning, while reranking serves as a critical precision amplifier in constrained retrieval environments.

Our experiments established a clear hierarchy of competence, with the proposed method consistently outperforming both Single-Hop and Standard Multi-Hop baselines, achieving a peak accuracy of roughly 60% on medium-difficulty bridge questions. We empirically validated a "Wide Net, Tight Filter" strategy, showing that a large initial fetch size ($N_{cand} = 30$) coupled with a Cross-Encoder Reranker effectively mitigates semantic drift by filtering out high-confidence distractors that often confuse standard vector search. Furthermore, our efficiency analysis revealed a distinct "Recall Ceiling," where blindly increasing the candidate pool beyond a certain threshold yield diminishing returns. This indicates that the primary failure mode for hard questions is not ranking error, but the fundamental failure of the embedding model to retrieve the correct document in the first place.

Looking forward, the performance degradation observed on "Hard" questions highlights the necessity for upstream improvements beyond simple reranking. Future research should prioritize integrating sparse retrieval (BM25 []) with dense embeddings to improve recall for rare entities often missed by semantic vector search. Additionally, implementing advanced query rewriting techniques such as HyDE [] could generate more diverse search queries, increasing the probability of retrieving the correct document during the initial fetch stage. Finally, exploring end-to-end optimization by fine-tuning the embedding model specifically for multi-hop reasoning tasks offers a promising avenue to align the vector space more closely with the logical structure of bridge questions.

References

[1] Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." Advances in neural information processing systems 33 (2020): 9459-9474.
[2] Gao, Yunfan, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. "Retrieval-augmented generation for large language models: A survey." arXiv preprint arXiv:2312.10997 2, no. 1 (2023).
[3] Guu, Kelvin, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. "Retrieval augmented language model pre-training." In International conference on machine learning, pp. 3929-3938. PMLR, 2020.
[4] Yang, Zhilin, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. "HotpotQA: A dataset for diverse, explainable multi-hop question answering." In Proceedings of the 2018 conference on empirical methods in natural language processing, pp. 2369-2380. 2018.
[5] Ho, Xanh, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. "Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps." arXiv preprint arXiv:2011.01060 (2020).
[6] Fang, Yuwei, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. "Hierarchical graph network for multi-hop question answering." In Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), pp. 8823-8838. 2020.
[7] Fu, Ruiliu, Han Wang, Xuejun Zhang, Jun Zhou, and Yonghong Yan. "Decomposing complex questions makes multi-hop QA easier and more interpretable." arXiv preprint arXiv:2110.13472 (2021).
[8] Zhou, Denny, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans et al. "Least-to-most prompting enables complex reasoning in large language models." arXiv preprint arXiv:2205.10625 (2022).
[9] Press, Ofir, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. "Measuring and narrowing the compositionality gap in language models." In Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 5687-5711. 2023.
[10] Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. "Chain-of-thought prompting elicits reasoning in large language models." Advances in neural information processing systems 35 (2022): 24824-24837.
[11] Karpukhin, Vladimir, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. "Dense Passage Retrieval for Open-Domain Question Answering." In EMNLP (1), pp. 6769-6781. 2020.
[12] Nguyen, Tri, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. "Ms marco: A human-generated machine reading comprehension dataset." (2016).
[13] Nogueira, Rodrigo, and Kyunghyun Cho. "Passage Re-ranking with BERT." arXiv preprint arXiv:1901.04085 (2019).
[14] Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks." arXiv preprint arXiv:1908.10084 (2019).
[15] Liu, Nelson F., Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. "Lost in the middle: How language models use long contexts." Transactions of the Association for Computational Linguistics 12 (2024): 157-173.
[16] Mallen, Alex, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. "When not to trust language models: Investigating effectiveness of

parametric and non-parametric memories." In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 9802-9822. 2023.
[17] Gao, Luyu, Xueguang Ma, Jimmy Lin, and Jamie Callan. "Precise zero-shot dense retrieval without relevance labels." In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1762-1777. 2023.
[18] Shen, Tao, Guodong Long, Xiubo Geng, Chongyang Tao, Tianyi Zhou, and Daxin Jiang. "Large language models are strong zero-shot retriever." arXiv preprint arXiv:2304.14233 (2023).
[19] Liu, Yang, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. "G-eval: NLG evaluation using gpt-4 with better human alignment." arXiv preprint arXiv:2303.16634 (2023).
[20] Zhu, Lianghui, Xinggang Wang, and Xinlong Wang. "Judgelm: Fine-tuned large language models are scalable judges." arXiv preprint arXiv:2310.17631 (2023).
[21] Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas et al. "Training compute-optimal large language models." arXiv preprint arXiv:2203.15556 (2022).
[22] Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. "Scaling laws for neural language models." arXiv preprint arXiv:2001.08361 (2020).
[23] Topsakal, Oguzhan, and Tahir Cetin Akinci. "Creating large language model applications utilizing langchain: A primer on developing llm apps fast." In International conference on applied engineering and natural sciences, vol. 1, no. 1, pp. 1050-1056. 2023.
[24] Jin, Jiajie, Yutao Zhu, Zhicheng Dou, Guanting Dong, Xinyu Yang, Chenghao Zhang, Tong Zhao, Zhao Yang, and Ji-Rong Wen. "Flashrag: A modular toolkit for efficient retrieval-augmented generation research." In Companion Proceedings of the ACM on Web Conference 2025, pp. 737-740. 2025.
[25] Kwon, Woosuk, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. "Efficient memory management for large language model serving with pagedattention." In Proceedings of the 29th symposium on operating systems principles, pp. 611-626. 2023.
[26] Douze, Matthijs, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. "The faiss library." IEEE Transactions on Big Data (2025).
[27] Borgeaud, Sebastian, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche et al. "Improving language models by retrieving from trillions of tokens." In International conference on machine learning, pp. 2206-2240. PMLR, 2022.
[28] Jiang, Zhengbao, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. "Active retrieval augmented generation." In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 7969-7992. 2023.

Authors

Hail Lim graduated from Michigan State University in Spring 2025 with a degree in Computer Science. His work focuses on AI, computer vision, and data engineering, including projects in RAG systems, YOLO-based wildlife monitoring, and robotics. He plans to pursue graduate studies in data science.

Toward Harmonized Global Governance of GMOs: A Comparative Analysis of Regulatory Fragmentation Across Kenya, the United States, and India

Junehyuk Kwak, St. Paul Preparatory Seoul, 50-11 Banpo-dong, Seocho District, Seoul, South Korea
Corresponding author email: junehyuk.kwak@gmail.com

Abstract

Genetically modified organisms (GMOs) now play an important role in global agriculture, but the rules that govern how they are developed, traded, and controlled remain separated. Most countries follow two main international systems, the WTO's trade-focused framework and the Cartagena Protocol's precautionary biosafety approach. Because these systems do not fully align, governments often face unclear or conflicting expectations when they try to regulate GM technologies. This paper examines these weaknesses through three case studies: Kenya's reversal of a long GMO ban during a climate-driven food crisis, the United States' repeated contamination incidents involving GM and non-GM crops, and India's dependence on Bt cotton, along with hindered regulatory progress and export contamination cases. Three case studies show that ununified governance increases suspicion, creates ecological and economic risks, and slows down the potential benefits that GMOs could provide for climate adaptation and food security. The challenges identified in the cases, including gene flow, resistance, and cross-border contamination, are transnational and cannot be addressed effectively by individual countries acting alone. For the reasons above, this paper argues that a unified international biosafety institution is needed to set minimum standards, coordinate responses to contamination, and provide clearer expectations around coexistence and liability. Such an institution would not remove national control but would help create a more predictable and coordinated global system for GMO governance.

Keywords

Genetically Modified Organisms (GMOs), Global GMO Governance, International Biosafety, Regulatory Fragmentation, Cross-border Contamination

Introduction

GMOs have been the center of recent biological advancements that change the nature of the organisms that humanity knows. The relatively rapid and recent nature of these developments, along with the complexities of how GMOs interact with their environment, also means that creating consensus on their use and developing effective standards of regulation remains challenging. While recognizing the benefits of GMOs, to ensure the safety of the environment, governments around the world have decided to create protocols to regulate GMOs. Two different international treaties for GMO regulations were established: The WTO's regulation and the Cartagena Protocol. This overlap in the regulation caused confusion and inefficiency in GMO trade and development. This paper's literature review explores the current difficulties created by reliance on the Cartagena Protocol and the WTO's Protocol, and the focus of the research is case studies on aspects of GMO regulation and the experiences of Kenya, the US, and India to inform understanding of how to improve the current system of regulation.

Thesis

This paper argues that the current global system for regulating GMOs is fragmented and does not match the cross-border nature of GMO-related risks. The WTO rules and the Cartagena Protocol create different and sometimes conflicting expectations for countries, which contributes to confusion, delays, and inconsistent policies. By analyzing the cases of Kenya, the United States, and India, this paper shows how these inconsistencies create real problems such as contamination incidents, sudden policy changes, and long periods with no regulatory progress. Because these issues repeat across different countries, the paper argues that a unified international body, or a much stronger coordinated framework, is needed to provide minimum standards, clearer monitoring systems, and consistent expectations for liability and coexistence. This type of institution would not remove national control but would help create a more stable and predictable global system.

Literature Review

This literature review will look into the inefficiency of the current regulation protocols of GMOs and bring up alternative proposals for regulating GMO trade. “Finalized in Nairobi in May 1992 and opened for signature at the United Nations Conference on Environment and Development (UNCED) in Rio de Janerio on 5 June 1992,” (CPB page 1), the Cartagena Protocol on Biosafety aims to ensure the protection of biodiversity and maximizing the “benefit from the potential that biotechnology has to offer.” The Cartagena protocol seeks to safeguard biological diversity and human health by regulating the international transfer, management, and use of living modified organisms produced through modern biotechnology. It emphasizes preventing potential harm, particularly in situations involving their movement across national borders.

Separate from the Cartagena Protocol, the WTO also passed its law for the regulation of GMO products. The Law on Genetically Modified Organisms has its focus on regulating the trading procedures of the GMOs and states that designated federal authority will evaluate whether the conditions have been met by reviewing the application and the report from the relevant organization, and will then decide whether to authorize the contained use, production release, or market placement of genetically modified organisms and their derived products.

On the Cartagena protocol, Robert Falkner and Aarti Gupta point out,“ the absence of a uniform global approach to GMO regulation, combined with disunity among leading agricultural trading partners in Europe has the effect of widening the space for autonomous decision-making in developing countries struggling with the challenges of domestic biosafety regulation.” (The Cartagena Protocol on Biosafety and Domestic Implementation) They also argue for a unified regulation of GMOs. Ultimately, however, Gupta and Falkner have contrasting views on the WTO’s regulation and support the Cartagena protocol for its allowance of more nationally tailored GMO regulation for each country.

Odong argues that the conflict between the WTO and the Cartagena Protocol “would be more pronounced” upon implementation of the Biosafety Protocol. States would necessarily have to put into practice their interpretation of the ambiguity created by the differences in the regulation. To avoid conflict between the two regimes, states could give priority to the trade-restrictive provisions of the Biosafety Protocol and further developments based on the protocol, making them exceptions to GATT/WTO rules and the chapeau to Article XX of the GATT. (Reconciling the Incongruence between the Cartagena Protocol on Biosafety and the GATT/WTO Rules). Robert suggests that GMOs should be regulated through the Sanitary and Phytosanitary Agreement: “As science is currently incapable of providing definite answers to the potential benefits and hazards of GMOs", priority should be given to the precautionary principle, especially with emerging international standards on GMOs, like the Codex Standards and the Cartagena Protocol (The ABC of GMOs, SPS & the WTO: An analysis of the application of the SPS Agreement to genetically modified organisms).

Researchers recognize that both the WTO and Cartagena's protocols offer valuable structure and guidance as foundation for the regulation of GMOs. Unregulated trading of GMOs poses significant potential threats that include but are not limited to the environment, human health, trust in international trade and biodiversity. Odong argues strongly for harmonization of WTO and GMO rules when regulating GMOs. As he argues, the inconsistent regulation of GMOs potentially causes inefficiency in the innovation and makes it harder to manifest the benefits that GMOs can undoubtedly offer in daily life. As for the problem of the application of GMOs, Robert argues that SPS's agreement states that regulations and restrictions should be restrictive as necessary, but this area is an important failure of both the WTO and the Cartagena Protocol. Scholars argue that the Cartagena Protocol's list of GMO registration is insufficient. Others argue that the WTO is insufficient in its containment policy and is focused too much in facilitating trade. Both views fail to offer substantive alternatives that address these issues. Currently, the proposed solutions focus on one aspect of the application of GMOs, which is its role in SDG and climate change. This paper will attempt to contribute to filling the gap in literature.

Methodology

This research uses legal analysis and three case studies to understand how the WTO rules and the Cartagena Protocol influence GMO regulation. The legal analysis identifies where the two treaties overlap, conflict, or leave important questions unanswered. The case studies show how these issues appear in real situations.

Kenya, the United States, and India were chosen because each country represents a different approach to GMOs. Kenya shows how policy can quickly change during a food security crisis. The United States shows the results of a permissive coexistence model that has experienced several contamination incidents. India shows how long delays in approving new GM crops create resistance problems and trade-related disputes. Together, these cases offer a broad look at how countries operate inside the fragmented global system.

The information used for the case studies comes from national laws, government statements, academic articles, and major contamination or policy events reported in the media.

Hypothesis

The main hypothesis is that the combination of WTO rules and the Cartagena Protocol creates a fragmented global system that does not match the transboundary risks associated with GMOs. Three expectations follow from this. First, different rules across countries increase the chances of contamination and resistance. Second, countries under pressure, such as drought or export problems, tend to make short-term and inconsistent decisions. Third, a unified international body with clearer standards could reduce these problems and make GMO governance more predictable. The case studies are used to examine whether these expectations match what happens in real situations.

Case Studies

1. Kenya's U-Turn on GMOs to Prevent Food Security Collapse

In November 2012, the Kenyan government imposed a moratorium on genetically modified organisms (GMOs), motivated by reports that GM maize caused cancer in rats (Joseph Maina, 2022). The Biosafety Act 2009 had established a regulatory framework for GMOs, but the cabinet's ban effectively prevented the commercialization of GM crops. However, in 2022, the Kenyan government was forced into a reversal on the commercialization of GMOs due to an unprecedented drought that threatened food security in the country. Although safety concerns had limited the adoption of biotechnology, the drought created environmental and population pressures that forced the government to allow the commercialization of GM crops after a decade of falling behind regional competitors.

During the moratorium, Kenyan scientists were limited to conducting confined trials and research on GMOs while lobbying the government for a change in the regulatory policy (Langat, 2022). During this period, South Africa, Nigeria, Sudan, and Burkina Faso moved forward with GM crop commercialization, creating a competitive advantage over Kenya in a relatively short space of time. With regional competitors benefiting from the development of GMOs and the perception that GMOs could increase food security during the drought, on October 3, 2022, Kenya's Cabinet formally lifted the decade-old GMO ban. In a special Cabinet meeting convened to address a worsening food crisis caused by climate change-induced environmental crisis, the Cabinet allowed the "open cultivation of genetically modified crops and the importation of food crops and animal feeds produced through biotechnology innovations" (Reuters, 2022). In practical terms, the Cabinet's 2022 decision nullified the 2012 moratorium and reinstated Kenya's pre-2012 regulatory regime under the Biosafety Act, 2009.

Before the ban on GMOs was lifted, due to the severe drought, four million Kenyans were facing acute food insecurity (Reuters, 2022). The national staple maize was in short supply, to the point where Kenya "consistently had an annual deficit of 10 million bags of the maize staple" (Duncan Miriri, 2022). The president, citing the need to stabilize food supply, allowed drought-tolerant and pest-resistant crops, and aligned the country's GM policy with scientific evidence. The policy change used GM crops to effectively counter the high levels of food insecurity that could have led to political and social instability.

This case proves the need for international GMO regulation, in which countries are free to use GM crops to address issues such as environmental and food insecurity. The case of Kenya shows that poorly informed domestic policy makers can damage development and would benefit from internationally agreed standards and data that can guide domestic decision-making. With a united international organization to regulate GMOs, countries can implement GM crops faster and with confidence to address potentially destabilizing threats such as climate change-related food insecurity.

With the practical need for GM crop adoption made clear by cases such as Kenya, the lack of standardized international regulation becomes a pressing issue, because policy differences can lead to cross-border contamination regardless of a state's desire to adopt genetic crops or not. The lack of regulation makes contamination more likely, as shown by the United States and India.

2. Cross-Contamination of GM Crops with Non-GM Crops in the US

Since the commercialization of GM crops in the United States, there have been high-profile cases of cross-contamination of various GM strains with non-GM strains. In the early 2000s, Starlink Corn was developed through genetic modification to create high levels of resistance to specific pests. The pesticidal protein carried potential allergy risks with human consumption, so the Environmental Protection Agency limited the use of the crop to animal feed. Tests on products in the human food supply chain, in products such as tacos, later found the presence of the protein, indicating the GM crop planted only for animal consumption had contaminated the non-GM form of the crop planted for human consumption (Jack Bobo, 2024). In another case of contamination, in 2006, a genetically modified rice developed by Liberty Link was found in the U.S. supply chain of rice for human consumption despite the company not receiving a federal license to grow commercial rice. (Plaintiff's Executive Committee, 2024) The owners of Liberty Link, Bayer CropScience, paid $750 million to settle claims against them (Plaintiff's Executive Committee, 2024).

Despite these high-profile cases, GM contamination of non-GM crops has been found extensively, including in rice, maize, soya, and rapeseed internationally (Becky Price & Janet Cotter, 2014). Unlike in the European Union, the United States has limited regulation on the prevention of cross-contamination of GM and non-GM crops. The U.S. has avoided specific legal provisions targeting producers. Instead, “segregation is achieved by those who pay the premium, with the implementation of locator maps, planting and buffer zones, third-party certification, cooperative exchanges…”(Rebecca Grumet, 2024). The approach in the U.S. can be defined as a “fence out” rule, which “imposes a segregation obligation on growers of organic, non-GM, and other crops that receive a premium” (Rebecca Grumet, 2024). This approach treats GM and non-GM crops equally, so there is no limit on the planting of approved GM crops or attempts by federal or local governments to limit contamination. Instead, “growers are obliged to ensure that GM pollen is excluded from their growing areas to guarantee that the crops they produce meet product quality requirements” (Rebecca Grumet, 2024).

In addition to the lack of control over cross-contamination of GM varieties created for non-human consumption with varieties created for human consumption, the extent of cross-contamination in the United States raises legal issues regarding the use of patented GM technologies. As the burden is placed on the non-GM farmer to prevent contamination, if contamination happens and the farmer, unaware, grows GM crops that have been patented, the farmer could be liable for use of the intellectual property, and biotechnology firms that develop, patent, and use GM crops have a track record of pursuing these claims in the courts (Food and Water Watch, 2015).

As GM and non-GM crops are traded extensively internationally, the potential for inadvertent cross-contamination and consumption is high when the regulatory system places the burden for preventing the contamination on the non-GM growers. As non-GM growers cannot efficiently trace the source of contamination, it is expensive and highly challenging to prevent it from happening. On the other hand, if the burden is placed on the GM crop grower, preventing “leaks” in a system designed to contain GM crops can be more easily enforced, as the technology will have a unique signature that can be traced to its source.

To achieve the higher level of safety and predictability associated with placing the burden on the GM producer, the U.S. could follow the E.U. and Japan in creating legislative frameworks that “mandate strict segregation measures for the cultivation of GM crops” (Rebecca Grumet, 2024). These frameworks created “systems for notification, segregation, labelling, delineation of planting areas, public registration, traceability, and compensation for damage” related to GM crops (Rebecca Grumet, 2024). The development of this framework in the E.U. has limited the use of certain types of GM crops that have a high risk for cross-contamination, such as insect-resistant maize (Rebecca Grumet, 2024).

Despite the constant threat of contamination of non-GM crops with GM varieties, the economic and environmental motivations for using GM remain powerful. The case of India shows how, despite concerns about cross-contamination and the development of disease resistance, when farmers have the choice, they overwhelmingly begin to rely on GM crops for their benefits.

In 2002, a GM cotton known as Bt cotton was introduced in India. Bt cotton effectively produces a pesticide, protecting it from pests known to cause significant damage to crop if not treated with environmentally damaging pesticides (Ian Plewis, 2021). A recent study shows that although there was an initial reduction in the use of pesticides after the introduction of Bt cotton, the GM modified plant has only continued to effectively repress one pest, while others have developed significant levels of resistance to the GM plant's pesticidal protein (K. R. Kranthi & Glenn Davis Stone, 2020). Furthermore, some researchers argue that the emergence of pesticide resistance caused by the introduction of Bt cotton means that cotton farmers are now spending more on pesticides than they were before the introduction of Bt cotton in 2002 (K. R. Kranthi & Glenn Davis Stone, 2020).

Since the commercialization of GM crops internationally, India's Genetic Engineering Appraisal Committee (GEAC) of the Ministry of Environment, Forest, and Climate Change has only authorized the use of Bt cotton. No other GM crops have been authorised for commercial use (PIB Delhi, 2024). Over 96% of cotton planted in India is now Bt cotton (PIB Delhi, 2024). Despite research showing that the introduction of Bt cotton may have created challenging pesticide resistance, in 2025, Indian researchers are demanding that the Indian government and Supreme Court develop more effective regulation and legislation to guide more widespread approval and use of GM crops. The principles for the development of new overarching regulation are listed as: "independent, evidence-based regulation; transparency and open data; structured public engagement; post-approval oversight; public investment in research; and a policy that protects organic and traditional farming, while giving farmers access to GM options without coercion." (Rohini Sreevathsa, 2025). The motivation behind the development of this legal framework will be a green revolution in agriculture that allows India to adapt effectively to rapid climate change (Rohini Sreevathsa, 2025).

The lack of regulation in India regarding commercialization and cross-contamination has led to disruption in supply chains, for example, rice. GM rice from field trials in India was found in rice exported to markets where no GM rice is permitted in products for human consumption, such as the E.U. (Claire Robinson, 2021). Due to such contamination events, significant concern has arisen over permission for the environmental release of GM mustard crops in India. Although the permission does not allow commercialization, the environmental release will create the potential for cross-contamination with non-GM mustard crops (GRAIN, 2023). Examples of this cross-contamination in the United States and with Indian GM rice have created the need for rapid development of a legal framework that preserves the choice of Indian farmers to use non-GM crops. Such a principle suggests that India will follow the EU and Japan in creating a legal system in which the burden is on GM crop producers to prevent contamination rather than following the U.S. approach in which non-GM producers face the burden of proving that their crops have not been contaminated.

Cross-Case Analysis

The case studies show the need for international cooperation on the creation of an international body to regulate and standardize certain aspects of the development and use of GMOs. Although market-driven competition for the use of GMOs has fueled innovation and value, the risks associated with cross-contamination with non-GMOs and disease/pest resistance require a comprehensive framework for international cooperation. The case studies show that although there are significant benefits to be gained from the development and use of GMOs, such as the initial reduction of pest threat in India and the ability to respond effectively to climate change-driven extreme weather events, such as drought in Kenya, the variation in regulation between jurisdictions risks creating mistrust and damage. In the case of India, the inadvertent export of GM crops to the EU market epitomizes how a lack of consistent regulation regarding cross-contamination can create tension and mistrust in trade relationships. The examples of cross-contamination within the US market show that financial damage can also be high when ineffective strategies for preventing cross-contamination are relied upon. As cross-contamination and the development of disease/pest resistance are inherently transnational, the only way to create an effective response to these downsides of the adoption of GMOs is through transnational cooperation. The US approach of "ring fencing" non-GMOs with the burden on non-GMO producers seems to ignore the real-world experience of cross-contamination and the potential effects of GMOs on the natural environment. One of the most significant concerns is the contamination of human food sources with GMOs that have been approved for non-human consumption, as shown by the examples in the US case study.

Recommendations

Based on the findings of this paper, the recommendation is for the creation of an international body that sets minimum standards for GM crop adoption. Membership of the body should be encouraged by offering shared expertise and other benefits within the group and cooperation to exclude non-members from trade and markets on the basis that their approach to GMOs does not meet the internationally agreed-upon standard. The body would first develop uniform rules that prevent confusion in the regulation process. Standardization of regulation will contribute to addressing issues such as cross-contamination and disease resistance. The body should also focus on an international standard for responses to inadvertent contamination to avoid export and market shocks. Such responses could include a clearly defined escalation of steps, with the final step being a comprehensive block of all exports when contamination is found and not addressed. The predictability of this type of approach will give companies clear guidelines and direction, which will create higher levels of stability in markets and confidence in consumers and governments.

Research Limitations

This study has several limitations that should be kept in mind for future studies. First of all, it lacks an adequate sample size to represent the whole world. In this paper, due to lack of time and resources, three case studies were analyzed: Kenya, the United States, and India. These countries represent different distinctive regulatory approaches, but they do not fully represent the other 190 countries. Thus, the findings should not be treated as complete or fully global.

Another limitation is that the research relies solely on publicly accessible sources such as national laws, academic articles, policy papers, and news reports. These sources provide useful information about rules and well-known events, but they sometimes do not reflect recent or unpublished policy discussions.

The third limitation is methodological limitation. The analysis is mostly qualitative, not quantitative. It focuses on legal frameworks, political decisions, and documented events rather than quantitative measurements of yields, scientific long-term impacts on nature, or the economic costs of contamination. As a result, the paper cannot make inarguable connections between regulatory choices and their environmental or market effects.

These limitations do not weaken the main suggestion that ununified international governance creates problems for effective GMO governance. However, they show that future research would benefit from a broader set of country cases, more quantitative data, and closer examination of how international policies are implemented in practice.

Conclusion

The case studies in this paper show that the main problem in GMO regulation is not whether countries use precaution or support trade. The real issue is that there is no unified system that governs the scientific and practical aspects of GMOs. Kenya's lifting its GMO ban during a severe drought, the United States' repeated contamination incidents, and India's pest control issues and slow approval times all demonstrate that gene flow and trait movement do not stop at the national borders. When countries follow different rules, the result is inconsistent policies and slow decision-making that does not keep up with advancements in biotechnology.

The comparison between the WTO system and the Cartagena Protocol shows clear limitations. Each framework focuses on a different priority, and neither provides complete regulations on contamination, coexistence, or the movement of unapproved GM traits through trade. Because of this disparity, farmers, traders, and governments face more uncertainties. They also increase the risk of market disruptions and damage credibility.

For these reasons, the paper argues that a consistent international biosafety institution is indeed needed. A global body with a clear mandate could help coordinate risk assessment, monitoring, and responses to contamination. It could also provide shared data, clearer standards, and more predictable expectations for liability. This would support innovative bio-tech advancements by giving countries and producers a more stable system while also improving safety for non-GM or organic product supporters.

Without stronger global coordination, countries will continue to rely on their own rules, and the same problems seen in the three case studies will continue to appear. As biotechnology spreads larger, the gap between reality and regulatory rules will only grow. Establishing a more coherent and concise global governance structure is truly essential for enabling GMOs to contribute to climate adaptation, food security, and sustainability in a responsible way.

References

1. Aarti Gupta and Robert Falkner. "The Cartagena Protocol on Biosafety and Domestic Implementation: Comparing Mexico, China and South Africa." Chatham House, Mar. 2006, p. 16, https://www.chathamhouse.org/sites/default/files/public/Research/Energy%2C%20Environment%20and%20Development/bp0306cartagena.pdf?utm_source=chatgpt.com.
2. Analysis of International Coexistence Management of Genetically Modified and Non-Genetically Modified Crops.https://www.mdpi.com/2223-7747/14/6/895#B13-plants-14-00895. Accessed 6 Aug. 2025.

3. “CredSpark.” CredSpark,https://app.credspark.com/assessments/copy-5ebc5436534e6-42ed6691-e693-4375-a04b-c4d7a602641d/assessment_responses/new. Accessed 30 July 2025.
4. Cultivation of Genetically Modified Crops.https://www.pib.gov.in/www.pib.gov.in/Pressreleaseshare.aspx?PRID=2042234. Accessed 6 Aug. 2025.
5. Cunningham, Robert. The ABC of GMOs, SPS & the WTO: An Analysis of the Application of The.
6. Editor2. Did Illegal GMO Rice Contamination Come from Indian GMO Field Trials? 18 Sept. 2021,https://www.gmwatch.org/en/main-menu/news-menu-title/archive/101-2021/19881-did-illegal-gmo-rice-contamination-come-from-indian-gmo-field-trials.
7. Entine, Jon, et al. “Regulatory Approaches for Genome Edited Agricultural Plants in Select Countries and Jurisdictions around the World.” Transgenic Research, vol. 30, no. 4, 2021, pp. 551–84. PubMed Central,https://doi.org/10.1007/s11248-021-00257-8.
8. Frontiers | Molecular Farming Navigates a Complex Regulatory Landscape.https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2024.1411943/full. Accessed 6 Aug. 2025.
9. Genetically Modified Cotton: How Has It Changed India?https://researchoutreach.org/articles/genetically-modified-cotton-how-changed-india/. Accessed 6 Aug. 2025.
10. GRAIN | GM Mustard in India: Thousands of Years of Cultural Heritage under Threat.https://grain.org/en/article/6944-gm-mustard-in-india-thousands-of-years-of-cultural-heritage-under-threat. Accessed 6 Aug. 2025.
11. “Kenya Approves GMOs after 10-Year Ban.” Alliance for Science,https://allianceforscience.org/blog/2022/10/kenya-approves-gmos-after-10-year-ban/. Accessed 2 Aug. 2025.
12. “Kenya Lifts Ban on Genetically Modified Crops in Response to Drought.” Reuters, 4 Oct. 2022. www.reuters.com,https://www.reuters.com/world/africa/kenya-lifts-ban-genetically-modified-crops-response-drought-2022-10-04/.
13. “Liberty Link Genetically Modified Rice.” Seeger Weiss LLP,https://www.seegerweiss.com/environmental-litigation/llrice-genetically-modified-rice/. Accessed 6 Aug. 2025.
14. Long-Term Impacts of Bt Cotton in India | Nature Plants.https://www.nature.com/articles/s41477-020-0615-5. Accessed 6 Aug. 2025.
15. Miriri, Duncan. “Kenya’s GMO Maize Push Sowing Trouble for Food Sector, Farmers Warn.” Reuters, 17 Dec. 2022. www.reuters.com,https://www.reuters.com/world/africa/kenyas-gmo-maize-push-sowing-trouble-food-sector-farmers-warn-2022-12-15/.
16. Mmbando, Gideon Sadikiel. “The Legal Aspect of the Current Use of Genetically Modified Organisms in Kenya, Tanzania, and Uganda.” GM Crops & Food, vol. 14, no. 1, pp. 1–12. PubMed Central,https://doi.org/10.1080/21645698.2023.2208999. Accessed 2 Aug. 2025.
17. October 2022, Anthony Langat // 12. “Kenya Lifts Ban on Genetically Modified Foods despite Strong Opposition.” Devex, 12 Oct. 2022,https://www.devex.com/news/sponsored/kenya-lifts-ban-on-genetically-modified-foods-despite-strong-opposition-104170.
18. Price, Becky, and Janet Cotter. “The GM Contamination Register: A Review of Recorded Contamination Incidents Associated with Genetically Modified Organisms (GMOs), 1997–2013.” International Journal of Food Contamination, vol. 1, no. 1, 1, Dec. 2014, pp. 1–13. foodsafetyandrisk.biomedcentral.com,https://doi.org/10.1186/s40550-014-0005-8.
19. “Reconciling the Incongruence between the Cartagena Protocol on Biosafety to the Convention on Biological Diversity and the GATT/WTO Rules.” ResearchGate.

www.researchgate.net,https://doi.org/10.1007/s10784-024-09649-7. Accessed 31 July 2025.

20. The Law on Genetically Modified Organisms. WTO, June 2020,https://www.wto.org/english/thewto_e/acc_e/cgr_e/wtacccgr9_leg_3.pdf.
21. United Nations Treaty Collection.https://treaties.un.org/pages/ViewDetails.aspx?src=TREATY&mtdsg_no=XXVII-8-a&chapter=27&clang=_en. Accessed 30 July 2025.
22. Weinzapfel, Daniel M. "The Economic Argument for Expanding GMO Regulation in America." Inquiries Journal, vol. 9, no. 02, 2017. www.inquiriesjournal.com,http://www.inquiriesjournal.com/articles/1537/the-economic-argument-for-expanding-gmo-regulation-in-america.

A Media Affordance Perspective on Digital Hoarding Behavioural Motivational Mechanisms Among Female Youth Users: A Case Study of Rednote

Xinyuan Wu, City University of Hong Kong
Corresponding author email: wuxinyuan2025@yeah.net

Abstract

As social media's penetration into users' lives continues to deepen, individuals have increasingly engaged in digital hoarding behaviours when confronted with vast amounts of information. However, this phenomenon remains underexplored in current communication research. This study therefore focuses on female youth users of the Rednote, adopting a theoretical framework of media affordance and integrating four affordance dimensions—production, social interaction, mobility, and imagination. Through in-depth interviews, walkthrough methods, netnography, and thematic analysis of textual data, the thesis investigates motivations of digital hoarding among female youth users.

The findings reveal that four affordances—production, social interaction, mobility, and imagination—are actualised through the interaction between the platform and young female users, fostering digital hoarding via diverse pathways. This paper provides insights for subsequent social media platforms to help young female users with digital hoarding anxiety balance the relationship between self, technology, and information, ultimately achieving healthy and efficient management of digital resources.

Keywords

Digital hoarding; Media affordances; Female youth users

Introduction

In recent years, the term "digital hamster" has gained popularity online. This phrase vividly describes individuals who, in the information-saturated cyberspace, tirelessly use functions like collecting, liking, and sharing to hoard "mental aliment". digital hamsters reflect the information anxiety and media survival anxiety users experience when confronted with overwhelming data. As China's information technology infrastructure accelerates and social media diversifies, media content which is predominately structured by user-generated content (UGC) has surged, intensifying hoarding behaviors among these "digital hamsters".

As a representative mainstream UGC platform in China, Rednote generates over 70 trillion note impressions daily, with more than 95% being UGC content. According to the platform's user profile characteristics, female users account for 71.98%, and youth users aged 18-35 constitute over 79%. Therefore, this study will focus on female youth users as the primary research subjects and research on how digital hoarding behavioral motivational mechanism's function, so that social media platforms can help young female users with digital hoarding anxiety achieve healthy and efficient management of digital resources.

Backgrounds

Digital hoarding

In 2015, the world's first literature analyzing digital hoarding patients from a psychiatric perspective was published. Bennekom et al. [25] authored proposed that digital hoarding, as a distinct subtype of hoarding disorder, occurs when the accumulation of digital files overwhelms an individual's capacity to organize them, leading to psychological distress and functional impairment. Digital hoarding behavior reflects complex social contexts and evolving times. The Web 2.0 era witnessed the rise of social media, sparking a surge in UGC. Wang et al. [38] defined digital hoarding in social media environments as: "The indiscriminate saving of data and unwillingness to delete it exhibited by users, driven by the explosive growth of data volume and reduced costs of data storage in social media." Tan et al. [24] further subdivided digital hoarding into the concept of "UGC digital hoarding," positing that UGC digital hoarding is a subset of digital hoarding specifically referring to UGC as the hoarded object. It represents a particular information behavior spontaneously exhibited by users during daily browsing or searching for information on UGC platforms. Users save, bookmark, or download UGC for reasons such as perceived utility, yet excessive hoarding and poor management hinder its reuse. Given the high degree of autonomy and diverse motivations behind users' digital hoarding on UGC platforms, Tan et al. [25] avoids terms like "indiscriminate" used by Wang et al. [38], instead defining digital hoarding from the perspective of users' active information filtering.

Compared to the Web 1.0 era where users could only passively receive information, digital hoarding has become more prevalent in the Web 2.0 and even Web 3.0 eras dominated by UGC. It exhibits new behavioral characteristics and generation mechanisms, underscoring the critical importance of focusing related research on social media platforms centered around UGC. Furthermore, while some scholars argue that digital hoarding does not cause severe psychological issues like physical hoarding, prolonged failure to organize accumulated digital content can still evolve into pathological digital hoarding disorder [37], then disrupting individuals' normal lives.

The development of media affordance

The term "affordance" originated from Gibson’s [7] research in ecological psychology is defined as: directly perceived substances, surfaces, objects, and places with the potential for action. Gale and Parchoma [19] noted that Gibson’s theory overlooks the influence of sociocultural factors on human perception and interpretation of the environment. With sociocultural development, mobile media has gradually emerged, and its unique communication characteristics have infused new connotations into affordance theory. Against this backdrop, Schrock [23] pioneered the term "communicative affordances." Communicative affordances are described "the interaction between subjective perceptions of utility and objective technical properties that alter communicative practices". Crucially, Schrock posits that the formation of perceived utility relates to individual goals, rather than the direct perception under "animal needs" proposed by Gibson. Nagy and Neff [17] incorporated emotion into affordance considerations, proposing imagined affordance. This approach further emphasizes users’ subjective perceptions and choices regarding technology from the aspects of individual psychology and social characteristics. In 2017, media affordance was introduced to China by Pan [18]. Pan [18] proposed that contemporary new media affordances can be categorized into production affordances, social affordances, and mobile affordances, corresponding respectively to the medium's support for content production, social interaction, and temporal-spatial freedom. Each affordance is further composed of several enabling forces. Pan Zhongdang's theoretical framework of media affordance demonstrates remarkable integration and inclusivity, encompassing numerous attributes and characteristics of new media. It serves as an innovative theoretical analytical tool for interpreting and evaluating new media [10].

Application of Media Affordance in Rednote Research

Chinese scholars have employed media affordance as a theoretical lens to examine various phenomena on Rednote since 2023. By using Pan Zhongdang's media affordance framework and incorporating Peter Nagy and Gina Neff's imagined affordance theory, Wen Ying [26] found that refined content on Rednote induces anxiety among female users. Zhang Yiwen and Yang Hongqi [36] contend that an imagined affordability relationship also exists between Rednote’s male user base and the platform. Despite Rednote’s female-dominated user base, the community supports male users in self-presentation. Their positive evaluations of the community's favorable atmosphere further contribute to the platform’s construction of a female-oriented community. Yan Wanying [32] posits that Rednote, functioning as a form of "social currency," not only facilitates interaction among community users but also enables non-regular users to engage in discussions through the topics it provides. This perspective aligns with Leidner’s "secondhand effect" theory [12]. Leidner [12] discovered that enterprise social media can motivate users who do not actively engage with them by transmitting through the behaviors and activities of active platform users, which is named "secondhand effect". Additionally, scholars adopting an "environment + relationship" perspective found that the construction of Rednote's platform society, the dominance of algorithmic logic, the drive of individual altruism, and the permeation of cultural traditions collectively shape adolescents' mutual aid practices on Rednote [33]. Overall, media affordance theory offers a multidimensional framework to deeply analyze the complex relationship between Rednote's diverse ecosystem and user behavior, providing robust theoretical support for studying digital hoarding phenomena within the platform.

Methods

This study aims to explore motivational factors of digital hoarding among female youth users on Rednote. A search of CNKI using keywords "Rednote" "digital hoarding" revealed only one article employing quantitative methods to investigate the influencing factors of digital hoarding behavior among Rednote users. Although this article introduced cultural identity and physical hoarding behavior as additional factors to construct an expanded model based on the Theory of Planned Behavior [4], it narrowly defined digital hoarding on Rednote as merely "bookmarking notes," overlooking the platform's diverse digital content and functions. Therefore, this study adopts a mixed-methods research approach with a qualitative focus. It collects data through in-depth interviews while employing methods such as walkthrough and ethnography, then using thematic analysis to dispose data so that this paper can achieve a profound interpretive understanding of how technical characteristics and female youth users interact on Rednote.

Data collection

Interview outline design. Analysing Rednote’s platform ecosystem and interaction pathways through walkthrough method, the researcher integrated research questions and theoretical foundations and designed an in-depth interview outline structured around five sections: "Basic information," "Platform Features and Interactions," "Understanding of Digital Hoarding," "Digital vs. Physical Hoarding," and "Digital Hoarding and Mental Health." Considering that bloggers, as both creators and consumers, embody this dual identity more distinctly than regular users, the interview outline was adjusted accordingly for this group. For platform operators, the researcher developed a separate outline tailored to their specific roles and responsibilities. The detailed in-depth interview outline is presented in Appendix 1.

Participants. This paper defines female youth users on Rednote as those aged 18-35 who exhibit digital hoarding habits on the platform, demonstrate proficiency in utilizing its various features, and possess an understanding of its operational mechanisms and cultural context. Participants' sampling split in two stages. In the initial research phase, interviewees were primarily selected through convenience sampling and snowball sampling. To understand digital hoarding behaviors across different educational backgrounds and occupations, the second phase employed purposive sampling to recruit and interview groups not reached in the first phase. Ultimately, the researcher interviewed 19 female youth users on Rednote exhibiting varying degrees of digital hoarding behavior, along with 3 Rednote operations staff. To protect partcipants' privacy, participants were assigned anonymized codes: ordinary users and bloggers were randomly numbered A1-A19, while operations staff were designated Y1-Y3.

Rednote's community. UGC on Rednote is also crucial data for examining digital hoarding. Collecting such data not only further ensures the reliability and validity but also enriches the overall research. UGC on Rednote primarily consists of two types: user-posted notes and comments under these notes. Although it is impossible to determine users' actual ages in highly anonymous online communities, the textual materials obtained can still fairly accurately reflect the phenomenon of digital hoarding on the platform. Therefore, the paper searched Rednote for keywords such as "digital hoarding," "digital hamster," and "collecting dust in bookmarks," limiting content publication within the past six months. Ultimately, a total of 21 text materials relevant to the research theme were collected.

Data Analysis

Using qualitative analysis software Nvivo 15, the paper conducted standardized analysis of 22 in-depth interview transcripts and 21 text materials collected from Rednote following the framework of "units—clusters—categories." [27]. This process yielded 41 units, 15 clusters, and ultimately formed 4 categories: "Digital Hoarding Behavior," "Motivators of Digital Hoarding," "Psychological and Behavioral Outcomes of Digital Hoarding," and "Guidance Strategies for Digital Hoarding."

Theme Saturation Test

Saturation testing is one of the key factors in conducting comprehensive qualitative research. Corbin and Strauss [5] define "saturation" as "Do not emerge new categories or related themes." Although research can never achieve complete saturation, "when the researcher believes a category provides sufficient depth and breadth of information to understand a phenomenon, and its relationships with other categories have been clarified, the sampling can be considered adequate and the research saturated." During the initial phase of interview subject sampling, the researcher constructed a preliminary thematic coding model based on textual data from 10 participants. To further enrich the coding hierarchy, the research optimized the interview outline around developed concepts. 12 users were then specifically recruited for a second round of interviews. Subsequently, the researcher collected text materials related to digital hoarding topics from Rednote to further test theoretical saturation. Analysis revealed that all collected text materials could be incorporated into the existing model, with no emergence of new categories, clusters, or units. Therefore, the study was deemed thematically saturated.

Motivations Behind Digital Hoarding

To further explore the underlying motivations behind users' digital hoarding behaviors, this study employed Nvivo 15's cluster analysis function. Using word similarity as the clustering criterion, it calculated the Pearson correlation coefficients between hierarchical codes associated with the category "Motivations for Digital Hoarding" within the thematic coding model. The coefficient values and their corresponding correlation strength criteria are as follows: 0.8–1 indicates extremely strong correlation, 0.6–0.8 indicates strong correlation, and 0.4–0.6 denotes weak correlation [29]. As shown in the correlation tree diagram (Figure 3-1), the motivations for digital hoarding exhibit strong or extremely strong correlations with factors related to production, social, imagined affordance, individual characteristics, and the tendency toward digital hoarding triggered by physical hoarding. This indicates that all six factors are significant drivers influencing users' digital hoarding behavior. Among these, mobile affordance exhibits the highest correlation coefficient, indicating it is the most central factor influencing digital hoarding motivations. Strong or extremely strong correlations also exist between all other clusters and units. Therefore, by referencing and deeply analyzing qualitative data materials, this chapter will examine how these factors facilitate female youth users' digital hoarding interactions with Rednote.

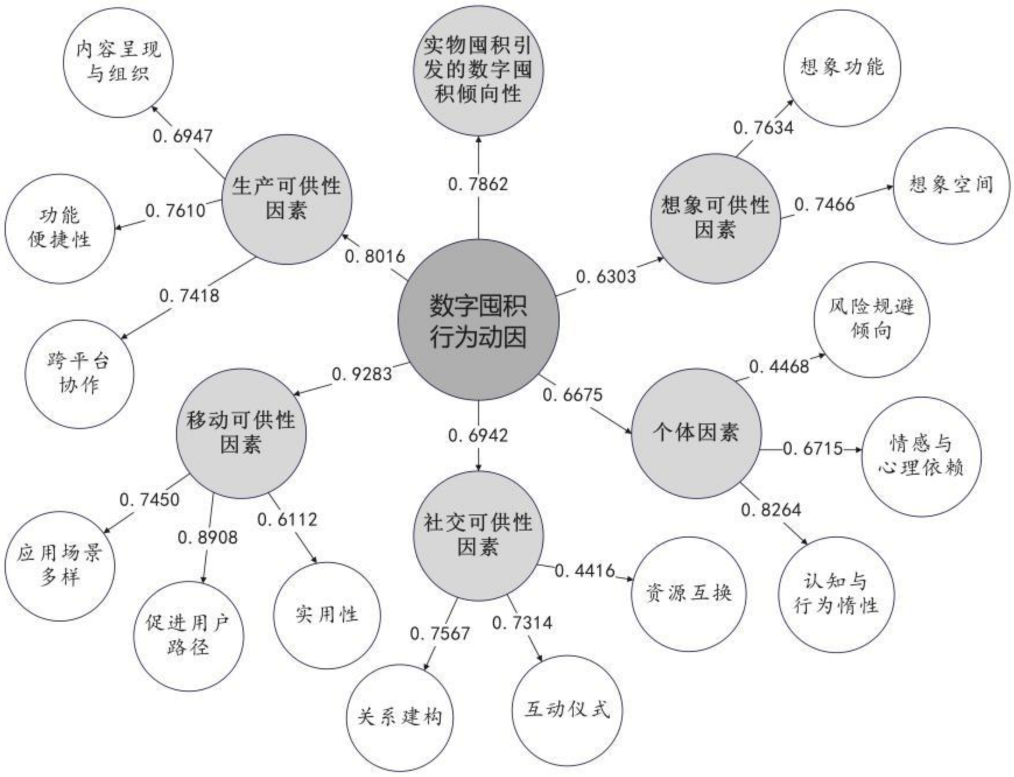


Figure 3-1

Production Affordance

Content presentation and organization. Rednote's content presentation and organization are divided into two units: content characteristics and update frequency. Both of them facilitate users' digital hoarding behavior. As China's largest sharing community, Rednote covers virtually every aspect of practical content—from shopping, fitness, and beauty to travel, work, and learning. It has become a "Google" for many women, "Its user base is massive, and no matter what kind of information you seek, countless people share it, which allowing you to find and save exactly what you want" (A9). However, the massive user base of social media inevitably generates vast amounts of repetitive, unstructured information. The platform's image-text-focused production model further accelerates content iteration, leading to information overload among users. Information overload occurs when an individual's processing capacity cannot keep pace with overwhelming information production, with social media being a primary source of this overload [3]. When users' information absorption pace lags behind the platform's production rate, digital hoarding behavior emerges.

"For tutorials on makeup or hairstyling, I might search again when I actually need them because these trends evolve rapidly. Searching brings up new posts, making previously saved content redundant." (A6)

Beyond the practicality and massiveness, the fragmented ambiguity and refined presentation of notes also fuel users' hoarding impulses. A key factor lies in many creators employing clickbait tactics and information cramming to chase traffic. Some interviewees mentioned that when faced with such information-heavy content, they often adopt a "save first, view later" approach. However, due to the vagueness of titles, the strategy can make it difficult for users to later retrieve specific notes by searching for precise keywords. With no other choice, they end up hoarding new notes.

"I saved a note about bubble tea before, but bubble tea was just one topic within that main note. Its title had nothing to do with bubble tea, so I couldn't find it no matter how hard I searched." (A13)

Additionally, Rednote primarily serves urban women in first and second-tier cities, whose content often exhibits refined characteristics. This stems from the inherent virtuality of social networks, which encourages individuals to present their best selves. "People share their best moments on social media, but I feel this behavior is more pronounced on Rednote. Especially for someone who initially didn't even know about islands—when you see the ocean in the Maldives, it's hard not to save it." (A16) Over time, this interplay between technology and users has cultivated a content ecosystem centered on "refined living." Content that is aesthetically pleasing and challenges perceptions often garners significant user attention.

Functional Convenience. Amid the wave of digital content consumption, platform functional design profoundly shapes user behavior patterns. Rednote's "favorite" mechanism is remarkably simple: users need only tap the favorite button to save notes to their default saves or created boards. Wilbur Schramm observed that "when individuals choose information pathways, all other things being equal, they tend to flow along the path of least resistance" [22]. To explain this phenomenon, he proposed the formula: Potential Reward / Effort Required = Probability of Choice. In the context of this study, potential reward refers to the likelihood of users discovering practical content within Rednote's vast information pool, while effort required denotes the difficulty audiences face in accessing useful information. When one-click actions like saving, following, or liking replace the mental effort of memorization, making information acquisition and utilization more effortless, users' probability of saving content significantly increases. They become less inclined to rely on their own judgment to filter information value, instead evolving into a conditioned reflex of digital hoarding.

"One-click saving, following, or liking is just too easy. You might unconsciously perform these actions many times a day without even realizing it." (A12)

"Because of the ‘saves’ feature, I might lose patience while reading. I might just skim through it the first time, get a rough idea that it might be useful in the future, and then save it." (A19)

Meanwhile, on Rednote where the lines between likes and saves blur, likes has become another way users hoard content. "Honestly, I think likes and saves serve the same purpose. Sometimes I just instinctively hit ‘save,’ other times ‘like.’ For me, whether it's a like or a save, it has only one meaning: making it easier to find later." (A3) From the platform designers’ perspective, the like function was originally intended to help users express agreement and resonance. However, interviews revealed that users often don’t consciously distinguish between liking and save. These buttons, combined with Rednote "Search my notes, collects and likes" feature, allow users to revisit content when needed. However, this inevitably leads to more scattered and disorganized digital content, increasing the difficulty of managing it later.

Cross-Platform Hoarding: Replicable and Associable. Copying as a human activity has permeated virtually every aspect of formation, exchange, and sharing within human civilization, playing a vital role in social interactions. Unlike the authoritative information and elite discourse disseminated by traditional media in the past, copying has evolved into various functional buttons on social platforms, becoming easier and more frequent to use than ever before [10]. Screenshots, as a technology enabling precise replication of screen content, offer immense convenience for users to acquire, save, and utilize digital resources. Research indicates that 98% of users engage in screenshotting nearly every day [1]. The built-in screenshot function in mobile operating systems makes this cross-platform copying behavior highly prevalent on Rednote. This phenomenon is not only a result of technological advancement but also reflects how individuals actively engage in "reproducing space" through media practices [8]. The de-authorization and decentralization of screenshots provide an entry ticket for nearly all users, which led them to develop a "screenshot habit." Local photo albums serve as either "transit stations" or "final destinations" for these screenshots, becoming another significant space where digital information gathers dust. Many interviewees mentioned cross-platform hoarding on Rednote, accumulating content like memes, wallpapers, online avatars, and user comments.

"I only clean out my album every long while, so now it's full of useless screenshots, avatars, and tons of meme." (A4)

"Like just the other day, I changed my phone wallpaper and searched through so many options. They’re still sitting on my album now. I tried each one, then forgot to delete them afterward—that happened a lot." (A8)

Beyond screen captures, respondent A6 noted: "I copy abstract captions on the spot to ‘annoy’ my friends, but I don't intentionally hoard them. Since I've copied them, they remain in my keyboard’s clipboard history—which feels like incidental hoarding." This demonstrates how Rednote integration with input method platforms enables users to engage in passive hoarding through copy-and-paste functionality.

Social affordance

Resources reciprocity. The term "reciprocity" originates from social exchanges between individuals motivated by the expectation of reward. Such exchanges encompass not only material benefits like money, compensation, or real estate, but also intangible gains derived from social relationships themselves—such as social recognition, a sense of belonging, gratitude, and affection [2]. In the social media era, mutual aid has evolved into a new form of online mutual assistance, gradually shifting from equal mutual benefit to asymmetric resource exchange. On Rednote, resource exchange extends beyond mere commodity transactions to encompass the conversion of "economic-relational" capital between users, information sharing, and experience exchange. For instance, Respondent A1 stated: "Since I also take unpaid interviews regularly, I want you all to follow me so I can secure more promotional opportunities." On one hand, A1 can accumulate followers by assisting others, converting virtual social metrics into tangible commercial opportunities. On the other hand, those aided by A1 complete their academic tasks. This demonstrates that Rednote users tend to downplay traditional notions of equal reciprocity in cyberspace, prioritizing tangible personal gains instead. This qualitative perspective enriches the quantitative findings of scholars like Zhang Fang, who concluded that the sense of reciprocal equality exerts the least influence on users' psychological states and behavioral intentions [34].

Furthermore, Rednote group chats serve as bond connecting shared interests and online connections, mutually reinforcing digital hoarding behaviors between group administrators and members through mutual aid. Ordinary user A7 mentioned joining numerous groups:

"There's a group for wearable nails, four shopping groups, and exam study groups... Recently, 'A Journey of Flowers' has been quite popular, so I joined a group where the admin shared screenshots of Liu Xiaoqing's autobiography for everyone to see." (A7)

Users with similar interests to A7 satisfy their information needs by joining various groups, while group admins expand their influence and maintain community authority by sharing information resources—a practice that also reflects their desire to accumulate followers and group members.

However, group admins aren't always the sole information disseminators; they can also be the ones get shared with. "On Rednote groups, people mainly post seal pictures they've seen on the platform or photos they've taken while traveling. If a post gets reposted multiple times, I know that content is popular. Then I can remix it, and it'll probably go viral (A14)." Seal meme creator A14 established a group chat named "Seal's Nest," attracting numerous seal enthusiasts to share seal-related content. These chat logs indirectly became a resource repository for A14 to hoard, guiding their further creation and traffic generation.

Relationship Construction. Social cognitive theory suggests that users' behavioral choices are often influenced by anticipated outcomes. Building on this, scholars have found that people's expected internet outcomes can shape their online behaviors, with social outcomes being a component of these anticipated results. Social outcomes encompass factors such as gaining support and respect, communicating with others, and maintaining relationships [11]. When engaging in digital hoarding, users not only fulfill their social needs within the virtual world but also serve real-world social interactions. By collecting and sharing content within Rednote or sharing across platforms, they provide topics and resources for real-life social activities. "Because when you forward things to your friends, it's like even when you're on vacation and not at school, it's a way to stay connected. Plus, sharing stuff can also strengthen your relationship" (A6). Over time, cross-platform sharing accumulates content that occupies increasing storage space. Yet users often hesitate to clear it, fearing deletion of crucial information they wish to retain. Thus, forwarding is also considered an intangible form of hoarding [35].

Beyond this, users develop differentiated sharing strategies based on content and context. Digital scenarios typically involve spatial environments, real-time user states, habitual behaviors, and social atmospheres [15]. Rednote provides users with abundant communication resources, enabling them to generate a "sense of social presence" and select appropriate content for sharing based on the scenario, thereby extending real-life social relationships [30].

"For a while, I saved a ton of meme stickers... I joined a group on Rednote back then and ended up saving over 600 of these stickers (see Figure 4-1), all manually added to my WeChat... For example, when chatting with someone who loves soccer, I'd send a soccer-themed one. They have these specific logos, like mineral water or NetEase Cloud Music ones. I'd use these to replace text—like if someone says, ‘you should drink some water,’ I'd immediately send a water-themed one, or if they mention ‘listening to music,’ I'd send the NetEase Cloud Music-packaged sticker. It makes the conversation more fun and fits the context better." (A7)

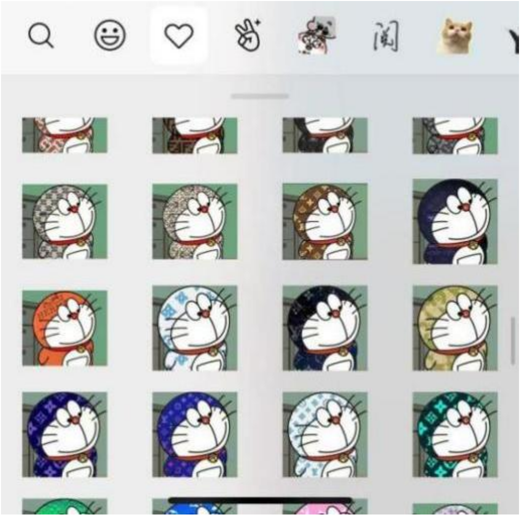


Figure 4-1 A7’s WeChat Chat Interface Emoji Bar

Additionally, Respondent A6 noted: "When hanging out with friends, if we're looking for places to go or spots to visit, or if I spot something fun, I'll forward it to share. Fun stuff usually gets shared right away, but food-related content might get saved for later so we can pick together." Users often have different real-time states when encountering various content types. Fun content is typically immediate and entertaining, requiring little deliberation or filtering. Users share it spontaneously based on their current mood, seeking feedback. Food-related content—like restaurant check-ins or dining recommendations—is more practical and decision-oriented. Choosing to save it before sharing allows users to gather more comprehensive and useful information, ensuring the shared content genuinely helps friends. This approach also earns them support and respect.

Interaction Ritual. In the era of individualization, the individual has become a fundamental unit within the social relationship system and an entity in the process of social action. Online social media further transforms individuals into nodes. Users not only appear as individuals but also form interconnected network structures between one another [31]. Comment sections serve as a crucial function providing connection opportunities. Along the entire content production chain, users accumulate not only original note content but also comments posted by users in comment sections—essentially hoarding these derivative texts.

"I might save a post because the comments are particularly insightful. Since posts are also places for sharing experiences, sometimes I find comments more insightful than the main post itself, or more useful to me, so I save them too" (A7).

"I just bought an air fryer a few days ago and searched online for lots of related stuff. Many netizens' comments below felt genuinely helpful, so I saved or liked them based on that" (A8).

The Rednote comment section isn't just a place for users to express feelings and interact; it's also a vital source of knowledge and experience. Users engage in discussions here because the notes themselves contain specific topics and content. Their exchanges resemble a "free-for-all debate" centered around a given theme. Through continuous questioning and answering, users foster the creation and circulation of knowledge and experience—often more practical than the original post itself.

Mobile Affordance

Space Eliminates Time: "Usefulness" on the Timeline. Peng examines the transformation of mobile accessibility under new media technologies from both temporal and spatial dimensions. Spatially, information consumption spaces have shifted from undifferentiated "wide-area spaces" to private and fluid "micro-spaces." While people still frequently enter diverse wide-area spaces for various activities, they have long since grown accustomed to creating entirely personal media spaces through mobile devices [20]. As a product of the mobile era, Rednote satisfies users' information-seeking needs across diverse scenarios while simultaneously increasing the potential for digital hoarding.

"I feel like I'm scrolling every hour—like when waiting for buses or subways, I'll open Rednote. I even scroll during class. It feels like I'm scrolling every second, except when I'm asleep." (A1)

"My commute is quite long—over two hours round trip—so I scroll on my phone. Since I don't like using Douyin, I just scroll through Rednote." (A3)

From a temporal perspective, the continuous evolution of social media has enabled users to construct personalized timelines where public and private information intertwine, and real-world existence blends with digital life [20]. Media technology continuously deconstruct and reconstruct people's perception of time. Fragmented browsing patterns make each individual's time unique, and the information they seek becomes increasingly personalized. When discussing reasons for hoarding notes, usefulness was a factor nearly every interviewee mentioned. In the historical flow of cyberspace, users can instantly retrieve digital information useful for the present or future via Rednote. This represents media's attempt to occupy all of an individual's media usage time—the phenomenon of "space eliminating time" [9].

"I've saved many nail art designs so I can show them directly at the salon. There are also clothes I want to buy when I return home, plus some Amsterdam delicacies I plan to try later." (A1)

Moreover, the dual discipline of social acceleration and time has gradually eroded people's autonomy. "Individuals no longer choose which new experiences to pursue; instead, the emergence of novelty drives them to rush toward it without reflection" [14]. Rednote user @SuiChuanguang remarked, "I rarely save posts after reflecting on them; most are saved on impulse—drawn in by a headline or a single line. Later, when I try to delete them, I can't overcome that 'usefulness' hurdle." Amidst the overwhelming tide of digital information, gauging usefulness has become a challenge for many users. Thus, rather than expending time and effort evaluating data's utility or value, it's more efficient to store as much digital content as possible in one's "information warehouse" within limited and accelerating timeframes. However, this approach also results in an extremely low "consumption rate" of digital information.

The synergistic effect of lightweight design and algorithms. Scholar Ding Yuanyuan, in her research on data journalism production strategies, points out that design minimalism—achieved through technical and content lightness—aligns with social platform users' mobile and fragmented news consumption habits. This enhances user experience while increasing the likelihood of content sharing [6]. Rednote's design minimalism manifests across both technical and content dimensions. On one hand, lightweight product recommendation links are deeply integrated with lifestyle content, reducing users' psychological sensitivity to commercial attributes and thereby triggering purchase-adding behavior.

"Rednote makes me accept that items I add to cart via links under posts might carry a premium. They might not be the cheapest online, but I don't obsess over price comparisons or question whether my purchase is cost-effective. Instead, it's more about the vibe of the product in the post you're viewing, or the feeling that buying it could grant you that lifestyle experience—it sparks an impulse to buy." (A15)

On the other hand, every note on Rednote features a minimalist card design with a "cover image + title" layout before opening. The relatively small number of notes displayed per page reduces cognitive load, making the mobile interface better suited for information consumption in scenarios with scarce attention. The "tap-based + swipe-based" browsing method allows mobile users to effortlessly transition to the next piece of content after watching a video, fostering uninterrupted browsing habits. Notably, the dissemination effectiveness achieved through technical and content streamlining relies heavily on further activation by the algorithmic system. By precisely delivering content aligned with user interests and implementing features like "Guess What You Want to Search," the algorithm continuously fuels users' hoarding tendencies.

"Among all the social media platforms I use, Rednote's algorithm feels the most responsive. It often seems to monitor my activity—or even after I view just one or two related posts—and immediately floods me with tons of similar content. This really fuels my hoarding behavior. Since it keeps pushing exactly the type of posts I want, I keep scrolling, and once I finish viewing, I save them." (A7)

Additionally, user A7 noted, "I don't intentionally click on posts from creators I follow. I usually just randomly scroll through the homepage. If it's someone you follow, the platform sometimes proactively recommends their content to you." During interviews, many respondents admitted they habitually browse the homepage recommendations, adopting a "whenever it happens" attitude toward content from their followed creators. This stems from Rednote's platform strategy to cultivate a browsing habit centered around its discovery page. Through deliberate interface design, the platform has conditioned users to prioritize this browsing pattern [16]. This also implies that even if some creators have large followings, their fans may not promptly notice newly posted content. The "Recommendations" page cannot cover all users' follows; instead, it recommends new creators to these fan groups. These users then follow the platform's algorithmic guidance to seek out and follow new high-quality creators, leading to the phenomenon of excessive accumulation of follows.

Imagine affordance: media support for the hoarding of digital memories

Memory forms the foundation of human cognition, self-construction, and self-reflection. As Rednote's primary mode of expression, image-text notes serve not only as channels for user sharing and interaction but also as memory carriers for the future, bearing the significance of historical reenactment [28]. Each Rednote note can accommodate up to eighteen images, accompanied by a title, lengthy text, and tags. Users can also edit notes at any time. This platform offers distinct expression methods and greater image capacity compared to WeChat Moments, thereby becoming a medium for users to reconstruct memories and preserve moments of forgetting.

Interviews revealed that many ordinary users view the "post notes" feature as "archiving memories." For instance, A5 stated, "I do it purely for my own records. I don't want to store them on cloud drives because accessing them is cumbersome. I'd rather keep them on a social platform I use regularly so I can browse them whenever I want." For food blogger A19, creating restaurant exploration notes isn't entirely altruistic: "About half the motivation comes from wanting to archive my dining experiences, while the other half is sharing my firsthand insights with a wider audience." They break free from the platform's pre-set narrative not to chase views, but out of concerns like "accidentally deleting my own photos" (A3), "losing track of my own content later" (A3), or "my own forgetfulness" (A2). Thus, the imaginative space co-created by users and the platform disrupts the linear timeline, allowing them to revisit the past at any moment. "When I opened my note about eating at Hamazushi last year, I noticed tiny details—like the service quality or how delicious it was—because it was my first time going with friends. Looking back now, it brings those memories flooding back" (A7). Through this cycle of "trace-recall-imagine," users evoke their own emotions, countering the fear of loss and forgetting.

Conclusions and Limitations

People are constantly acquiring and constantly forgetting. The evolution of digital existence has transformed individuals into "intermediary media" [21]. Individuals acquire, store, and forward knowledge and information through social media platforms, yet often remain stuck at the superficial level of "acquisition" rather than deep ‘absorption’ due to the mindset of "hoarding first, viewing later." This article integrates domestic and international theories of media affordance and imagined affordance, moving beyond previous platform-agnostic research to focus on digital hoarding behavior among young female users on Rednote. At the level of production affordance, Rednote's content production frequency and characteristics cause user information overload, fostering digital hoarding. Users then exacerbate this behavior through "hoard-first, view-later" practices. The convenience of "one-click" functions and copy-paste technology enables users to actively or passively hoard increasing amounts of digital content. At the social affordance level, the asymmetric mutual aid dynamics of the social media era incentivize hoarding. Through interactions like comments and group chats—primarily "weak ties"—users accumulate life experiences and creative inspiration on Rednote, while maintaining real-world social connections via cross-platform hoarding. Regarding mobile affordances, Rednote creates a "space-erases-time" media environment. Its lightweight content and technical design, combined with algorithmic recommendation systems, collectively facilitate users' hoarding pathways. At the level of imagined affordances, while the platform provides digital memory support, users transcend the platform's pre-set "technological scripts" to hoard their own emotional memories.

While this study offers innovative insights into young female users' digital hoarding behaviors and motivations, limitations remain. First, the research methodology primarily relies on semi-structured interviews and online field surveys. Although these methods can deeply explore user motivations, they still suffer from issues such as over-reliance on interviewers' subjective perceptions and insufficient observation duration. Future research could adopt a mixed-methods approach combining quantitative and qualitative data to develop a more suitable digital hoarding scale for social media contexts. This would allow broader exploration of hoarding behaviors across different groups. Increasing the frequency of follow-up interviews could also yield richer dynamic data for cross-validation. Second, the sample selection has limitations. Although 22 young female users were recruited through snowball sampling and online recruitment, these subjects were concentrated within specific gender, age, and occupational ranges, failing to represent all users. Future research should broaden the sample scope to enhance the generalizability of findings, thereby guiding whether platforms truly need to regulate users' digital hoarding behaviors and to what extent such regulation should be implemented. After all, "every small change made by the platform affects the lives and consumption experiences of hundreds of millions of users" (Y3). Finally, this study primarily examined digital hoarding on Rednote through the lens of media affordance theory, failing to fully leverage other perspectives to consider digital hoarding phenomena on Rednote and other social media platforms. Future research could explore the similarities and differences in digital hoarding across various social media platforms using different theoretical frameworks. It could also investigate how different platforms influence digital hoarding behaviors among users of various age groups. By examining both commonalities and differences, such research could explore how to enhance users' digital literacy and establish healthy, efficient digital lifestyles.

References

[1]. Aihaiti, R., Zheng, X., Li, S., & Fan, H. (2025). Construction of user screenshot behavior model in social media environment: Based on grounded theory. Document,Information & Knowledge, 42(1), 101–112.

[2]. Blau, P. M. (1964). Exchange and power in social life. Wiley.

[3]. Cheng, H., Yu, H., & Jiang, X. (2023). Antecedents and consequences of information overload of social media in major public health emergencies: A model. Information Science, 41(3), 45–56.

[4]. Cheng, S., Ruan, J., & Deng, X. (2024). Research on influencing factors of users' digital hoarding behavior on user-generated content platforms: Taking xiaohongshu as an example. Library and Information Service, 68(4), 58–69.

[5]. Corbin, J., & Strauss, A. (2015). Basics of qualitative research. SAGE Publications.

[6]. Ding, Y., & Zhang, C. (2019). Sharing as communication: A study on the socialized production strategies of data journalism. China Publishing Journal, 21, 5–9.

[7]. Gibson, J. J. (1979). The ecological approach to visual perception. Psychology Press.

[8]. Han, C., & Guo, C. (2023). A study on the spatial turn in media practices based on screenshot interfaces. Jiangxi Social Sciences, 43(3), 179–187.

[9]. Jiang, X., & Zhao, W. (2016). The evolution and integration of media's concepts of time and space in the post-internet era. Social Science Front, 11, 154–160.

[10]. Jing, Y., & Shen, J. (2019). The incorporation and expansion of the concept of new media affordances. Contemporary Communication, 1, 92–95.

[11]. LaRose, R., & Eastin, M. S. (2004). A social cognitive theory of internet uses and gratifications: Toward a new model of media attendance. Journal of Broadcasting & Electronic Media, 48(3), 358–377. https://doi.org/10.1207/s15506878jobem4803_2

[12]. Leidner, D. E., Gonzalez, E., & Koch, H. (2018). An affordance perspective of enterprise social media and organizational socialization. The Journal of Strategic Information Systems, 27(2), 117–138. https://doi.org/10.1016/j.jsis.2018.03.003

[13]. Li, X. (2019). The origin and rationality of "reproduction": A study of benjamin's communication thought [Master's thesis, Fujian Normal University].

[14]. Lian, S., & Deng, D. (2020). "Myths" of time and modernity: A critical study of communication from the perspective of social acceleration theory. Modern

Communication (Journal of Communication University of China), 42(6), 37–42.
[15]. Liao, W. (2023). The impact of metaverse on the scenario innovation of television programs. Nanfang Media Research, 03, 64–69.
[16]. Liu, Y., & Wang, W. (2022). Marking whose life: Scripts and digital gig workers on the xiaohongshu platform. Beijing Culture Creativity, 1, 79–89.
[17]. Nagy, P., & Neff, G. (2015). Imagined affordance: Reconstructing a keyword for communication theory. Social media + Society, 1(2), 1–9. https://doi.org/10.1177/2056305115603385
[18]. Pan, Z., & Liu, Y. (2017). What counts as "new"? Power traps in the discourse of "new media" and researchers' theoretical self-reflection — an interview with professor pan zhongdang. Journalism & Communication Review, 1, 2–19.
[19].Parchoma, G. (2014). The contested ontology of affordances: Implications for researching technological affordances for collaborative knowledge production. Computers in Human Behavior, 37, 360–368. https://doi.org/10.1016/j.chb.2012.05.028
[20]. Peng, L. (2022). Changes in communication affordances and their impacts under new media technologies. Modern Publishing, 6, 60–73.
[21]. Peng, L. (2024). The evolution of mediated survival in the 30 years of internet development. Editorial Friend, 10, 5-14+36.
[22]. Schramm, W., & William Earl Porter. (1982). Men, women, messages, and media. Harpercollins College Division.
[23]. Schrock, A. R. (2015). Communicative affordances of mobile media: Portability, availability, locatability, and multimediality. International Journal of Communication, 9(1), 1229–1246.
[24]. Tan, C., Zou, Y., Wang, Y., & Wang, X. (2025). Research on the formation mechanism of UGC digital hoarding behavior of social media users. Library Tribune, 45(3), 130–140.
[25]. van Bennekom, M. J., Blom, R. M., Vulink, N., & Denys, D. (2015). A case of digital hoarding. BMJ Case Reports, bcr2015210814. https://doi.org/10.1136/bcr-2015-210814
[26]. Wen, Y. (2023). Effect of xiaohongshu use on female users' anxiety: A perspective of the media affordances [Master's thesis, Southwest Jiaotong University].
[27]. Wu, X. (2023). Shaping visitors' word of mouth towards immersive performing art： The role of interactive experience [Doctoral Dissertation, Southwestern University of Finance and Economics].
[28]. Wu, X. (2024). Research on memory construction of personal digital archives— A case study of the annual usage report of social media APP [Master's thesis, Shanxi University].
[29]. Wu, X., & Tan, X. (2022). Time for space: The "time-spacial adaptation" mechanism of grassroots governance policy innovation—investigation on the residence property service innovation in chengdu. Journal of Public Management, 19(3), 123-135+174.
[30]. Wu, Z. (2023). The construction and transformation of social media relationships from the perspective of media situation theory. West China Broadcasting TV, 44(12), 38–40.
[31]. Xiong, Y. (2024). Swing, performance and social contact：social media platform practice of young users in the polymedia environment [Master's thesis, Hubei University].
[32]. Yan, W. (2024). A Study on the "Search Engineization" Phenomenon of Xiaohongshu from the Perspective of Media Affordances Theory. New Media Research, 10(16), 76-78+97.
[33]. Yan, W., & Cao, W. (2024). An analysis of the influencing factors of adolescents' online prosocial behavior from the perspective of affordances — taking xiaohongshu as a case study. West China Broadcasting TV, 45(10), 17–20.
[34]. Zhang, F., & Liu, H. (2022). Research on the motivation of continuous participation

in online mutual help from the perspective of reciprocity. Journal of Fujian Normal University (Philosophy and Social Sciences Edition), 6, 107-118+171.
[35]. Zhang, X. (2024). A study on young adults' digital hoarding behavior from the perspective of affordances [master's thesis, Heilongjiang University].
[36]. Zhang, Y., & Yang, H. (2023). The construction of female-oriented groups by xiaohongshu from the perspective of media affordances. Public Communication of Science & Technology, 15(14), 101–104.
[37]. Zhao, D. (2024). Personal digital hoarding behaviors in the big data environment:a grounded theory study. Journal of Library Science in China, 50(1), 96–114. https://doi.org/10.13530/j.cnki.jlis.2024008
[38]. Wang, L., Du, T., & Zhu, H. (2022). The formation mechanism of data hoarding behavior in social media context: Taking college students as an example. Information Studies: Theory & Application, 45(1), 22–29.

Leveraging Machine Learning for Stroke Prediction: An Empirical Study on Clinical and Behavioral Risk Factors

Jungmin Lim, University of California, Irvine, 260 Aldrich Hall, Irvine, CA 92697
Corresponding author email: Jungminlim1015@gmail.com

Abstract
This study investigates the applications of machine learning techniques for predicting the stroke risks using clinical, behavioral and demographic features. Multiple classification models were evaluated, and the random forest classifier achieved the highest performance, with a recall rate for stroke of 98% and an AUC of 0.98. Feature importance analysis showed that age, average glucose level, and BMI are the most influential predictors. From an operational perspective, integrating predicting modelling into healthcare systems can facilitate early risk detection and support personalized care strategies.

Keywords
Stroke, Healthcare, Machine Learning, Random Forest

Introduction
Based on the World Stroke Organization's estimates, approximately 15 million people suffer from a stroke each year, and about 5 million die from stroke-related reasons (World Health Organization, 2021). Stroke is one of the leading causes of death and disability. It not only affects those who are experiencing it but also their families and wider society (Edmans et al., 2010), and it can occur in anyone at any age (Elloker & Rhoda, 2018). Therefore, a deeper understanding of the stroke mechanisms and establishing effective risk stratification strategies are urgently needed for both primary and secondary prevention. Managing modifiable risk factors could help prevent nearly half of all strokes among individuals at high risk (Brainin et al., 2018).Machine Learning (ML) has been rapidly developed and implemented across disciplines and is becoming a transformative force in healthcare research and practice (Dritsas & Trigka, 2022). By leveraging algorithms that are capable of capturing complex, and often hidden relationships among diverse clinical, demographic and physiological variables from many origins including patients' history, imaging and biomarkers (Singh et al., 2025), ML enables more accurate prediction and decision-making than traditional statistical methods. One application of ML is in precision medicine, where ML models are developed to identify the most effective treatment strategies based on an individual's unique conditions (Lee et al., 2018). Machine learning applications in healthcare have expanded rapidly. In the context of stroke, ML techniques have been increasingly used to identify and predict risk factors that contribute to stroke occurrence, recurrence, and recovery outcomes. Predictive models using logistic regression, random forests, gradient boosting, and neural networks have shown strong performance in recognizing key determinants such as age, hypertension, diabetes, smoking, and heart disease (Ahammad, 2022; Hassan et al., 2024; Khosla et al., 2010).Identifying these factors not only facilitates early intervention and resource allocation but also provides data-driven guidance for healthcare systems to design more efficient prevention and management strategies.

Backgrounds

Stroke is influenced by a complex interplay of nonmodifiable and modifiable risk factors. Nonmodifiable factors include age, sex, race–ethnicity, and genetics. The incidence of stroke doubles with every decade after age 55, and while the mean age of ischemic stroke remains around 69 years, recent evidence shows a rise among adults aged 20–54 years, increasing from 12.9% in 1993/1994 to 18.6% in 2005 (George et al., 2011; Kissela et al., 2012; Roger et al., 2012; Van Asch et al., 2010). The relationship between sex and stroke risk is age dependent: younger women face similar or slightly higher risk than men—likely due to pregnancy, hormonal contraception, and postpartum factors—whereas men have a higher risk in later life (Asplund et al., 2009; Kapral et al., 2005; Reeves et al., 2009). Racial and ethnic disparities are also pronounced. Black Americans experience twice the incidence and higher mortality rates than White populations (Cruz-Flores et al., 2011; Gillum, 1999b), with similar elevations reported among Hispanic/Latino and American Indian groups (Kleindorfer et al., 2006). These inequalities are attributed in part to higher prevalence of hypertension, obesity, and diabetes (Giles et al., 1995; Gillum, 1999a), but social determinants such as healthcare access, language, and nativity also play critical roles (Howard et al., 2011; Joubert et al., 2008). Genetic predisposition further contributes to stroke risk: a positive parental or family history increases the likelihood of stroke, with genetic effects varying across age, sex, and ethnicity (Seshadri et al., 2010).

Modifiable risk factors provide key targets for prevention. Hypertension remains the most significant, accounting for roughly 54% of stroke population attributable risk in the INTERSTROKE study (Donnell et al., 2010). The risk of stroke rises progressively with blood pressure, even below the hypertensive threshold (Vasan et al., 2002). Diabetes mellitus doubles stroke risk and accounts for about 20% of deaths in diabetic patients, with longer duration of diabetes further increasing risk (Banerjee et al., 2012; Sui et al., 2011). Other major modifiable factors include atrial fibrillation and atrial cardiopathy (Yiin et al., 2014), dyslipidemia (particularly high total cholesterol and low HDL levels) (Horenstein et al., 2002), sedentary lifestyle, poor diet, obesity, and metabolic syndrome (Zhou et al., 2007), as well as alcohol consumption, illicit drug use, and cigarette smoking—the latter nearly doubling stroke risk and contributing to 15% of stroke deaths annually (Kuo et al., 2013). Emerging evidence also links inflammation, infection, and air pollution exposure to increased stroke incidence (Kaptoge et al., 2010). Together, these epidemiological findings underscore that both traditional and novel risk factors—spanning biological, behavioral, and environmental domains—must be integrated into predictive models and prevention strategies.

Machine learning (ML) has emerged as a powerful analytical framework for predicting stroke risk by leveraging large-scale clinical and behavioral datasets. Numerous studies have applied classical ML algorithms—such as Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB), and K-Nearest Neighbours (KNN)—to identify individuals at high risk of stroke, achieving accuracies ranging from 82% to 96% (Kokkotis et al., 2022; Sirsat et al., 2020; Wu & Fang, 2020). Among these, tree-based ensemble methods such as RF and Gradient Boosting consistently outperform traditional classifiers in accuracy, precision, and AUC performance. For instance, several studies reported RF models achieving 95–96% accuracy (Geethanjali et al., 2021; Tazin et al., 2021), while NB and LR models achieved competitive but slightly lower results (82–86%) (Geethanjali et al., 2021). These results highlight ML's potential to capture nonlinear relationships between stroke risk and clinical features such as age, hypertension, heart disease, glucose level, and smoking status (Sailasya & Kumari, 2021).

Building on these foundational methods, researchers have explored advanced and hybrid approaches to enhance predictive performance. Shanthi et al.(2009) applied an Artificial Neural Network (ANN) to predict thromboembolic strokes, reaching 89% accuracy. Similarly, Ahmed et al. (2019) achieved 90% accuracy using ML algorithms on the Apache Spark platform, while Tazin et al. (2021) improved accuracy to 95% after applying normalization and feature-ranking procedures. Other hybrid models—such as the Minimal Genetic Folding (MGF) algorithm (Mezher, 2022) and the RXLM ensemble combining RF, XGBoost, and LightGBM (Alruily et al., 2023)—further advanced predictive capacity, achieving 83% and 96.3% accuracy, respectively. To address data imbalance and enhance generalizability, many studies incorporated techniques such as SMOTE oversampling, feature selection, and outlier control (Sowjanya & Mrudula, 2023; Wongvorachan et al., 2023). Some even implemented real-time or cloud-based prediction tools that can collect user data and deliver early stroke warnings with 96% accuracy (Islam et al., 2021).

While progress in predictive modeling is significant, key methodological and practical challenges remain. Many existing studies rely on relatively small or imbalanced datasets, or focus on a limited number of attributes, constraining model robustness (Chen, 2023; Nijman et al., 2022; Paul et al., 2022). Furthermore, the high-performing ensemble and neural network models often function as “black boxes,” limiting interpretability and hindering clinical adoption. Another limitation lies in the lack of standardized evaluation metrics and external validation, which restricts comparability across studies. Therefore, the literature increasingly calls for the development of explainable ML frameworks, integration of diverse clinical and behavioral features, and comprehensive benchmarking on larger datasets. Such efforts are critical to ensure that predictive analytics can move beyond model optimization toward actionable, interpretable tools that support early stroke prevention and healthcare decision- making.

Methods

Our research utilized the publicly available Stroke Prediction Dataset from Kaggle (Stroke Prediction Dataset, 2025). From this dataset, we included only participants having no missing values, resulting in a total sample size of 4909 individuals. The dataset contains 10 predictor variables and one binary outcome variable indicating whether the participant has ever experienced a stroke. The predictors are defined as follows: Age (in years), Gender, Diagnosed hypertension, Heart Disease, Ever Married, Work Type (5 categories: private, self- employed, government job, never worked and children), Residence Type (urban, rural), Average Glucose Level (mg/dL), Body Mass Index (BMI) (kg/m^2), and Smoking Status (three categories: currently smokes, never smoked, and formerly smoked). The outcome variable, Stroke, represents whether the participant has previously suffered a stroke. Among these variables, age, average glucose level, and BMI are continuous, while the remaining features are categorical. We normalized the continuous variables and performed one-hot encoding for the categorical variables. To address the class imbalance between stroke and non-stroke cases in subsequent analyses, we applied the Synthetic Minority Oversampling Technique (SMOTE) (Maldonado et al., 2019), which synthetically augments the minority (stroke) class to achieve a balanced dataset for model training.

Machine Learning Models

Random Forest Classifier

The Random Forest (RF) algorithm is an ensemble learning method that constructs multiple decision trees and aggregates their predictions to improve generalization and classification performance. Given a training dataset:

$$D = \{ (x_i, y_i)\}_{i=1}^{n}$$

where $x_i \in \mathbb{R}^p$ denotes the feature vector with p predictors and $y_i \in \{0,1\}$ represents the binary class label (stroke or non-stroke). The RF algorithm performs the following steps:

1. Bootstrap sampling: Draw T bootstrap samples from the training set.
2. Tree growth: For each sample, grow an unpruned classification tree. At each node, a random subset of $m < p$ features is selected, and the best split among these m features is chosen to minimize impurity (e.g., Gini index or entropy).
3. Aggregation: Each tree $h_t(x)$ provides a class prediction. The final prediction of the forest is obtained by majority voting across all trees:

$$\hat{h}_{RF(x)} = mode\{ h_{t(x)}\}_{t=1}^{T}$$

4. This ensemble approach reduces variance and mitigates overfitting by combining multiple decorrelated classifiers.

k-Nearest Neighbor (kNN) Classifier

The k-Nearest Neighbor (kNN) algorithm is a non-parametric, instance-based learning method that classifies a new observation based on the majority label among its nearest neighbors in the training set. For any two data points x_i and x_j, the distance function is defined as:

$$d(x_i, x_j) = \left|\left|x_i - x_j\right|\right| = \sqrt{\Sigma_{p=1}^{P}\left(x_{ip} - x_{jp}\right)^2}$$

Given a new observation x_0, the classifier identifies its k nearest neighbors, denoted $N_k(x_0)$, and assigns the most frequent class label among them:

$$\hat{h}_{kNN(x^0)} = mode\left\{ y_i : x_i \in N_{k(x^0)}\right\}$$

For binary classification problems where $y_i \in \{-1, +1\}$, the decision rule can equivalently be written as:

$$\hat{h}_{kNN(x^0)} = sign\left(\left(\frac{1}{k}\right)\Sigma_{x_i \in N_{k(x^0)}} y_i\right)$$

The hyperparameter k controls the bias–variance trade-off: smaller k values lead to lower bias but higher variance, while larger k values produce smoother decision boundaries with higher bias.

Logistic Regression

The Logistic Regression (LR) algorithm is a statistical learning method used for binary classification problems. It models the conditional probability of the dependent variable $y_i \in$
$\{0,1\}$ given the predictors $x_i \in \mathbb{R}^p$ using the logistic (sigmoid) function. The model assumes a linear relationship between the predictors and the log-odds of the probability of the positive class. The logistic regression function is defined as:

$$P(x_i) = \frac{1}{\left(1 + e^{\{-(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})\}}\right)}$$

The logistic regression model estimates the coefficients $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ by maximizing the log-likelihood function $\ell(\beta)$:

$$l(\beta) = \Sigma_{\{i=1\}}^{\{n\}}[y_i \, log \, log \, (\pi_i) \, + (1 - y_i) \, log \, log \, (1 - \pi_i) \,], \;\; \pi_i = P(x_i)$$

The fitted probabilities can then be used for classification, where an observation is predicted as stroke-positive if the estimated probability exceeds 0.5. This model provides a simple, interpretable baseline for binary classification, assuming a linear relationship between predictors and the log-odds of the outcome.

Model Evaluation Metrics

Under the evaluation process of the considered machine learning (ML) models, several performance metrics were recorded. In the current analysis, we focus on the most widely used measures in related literature (Hossin & Sulaiman, 2015):

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1 = 2 \, \frac{(Precision * Recall)}{Precision + Recall}$$

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

$$AUC = \int_0^1 TPR(FPR) \, d(FPR)$$

Here, True Positives (TP) represent the number of participants who experienced a stroke and were correctly identified by the model as stroke cases. True Negatives (TN) denote the number of participants who did not experience a stroke and were correctly predicted as non-stroke cases. False Positives (FP) correspond to the number of participants who were incorrectly classified as having a stroke when they actually did not. And False Negatives (FN) refer to the participants who had a stroke but were mistakenly predicted as non-stroke.

From these quantities, we can derive two rates. True Positive Rate (TPR), also known as Recall or Sensitivity, quantifies the model's ability to correctly identify stroke cases and is computed as

$$TPR = \frac{TP}{TP + FN}$$

False Positive Rate (FPR) measures the proportion of non-stroke participants incorrectly classified as stroke and is defined as

$$FPR = \frac{FP}{FP + TN}$$

TPR and FPR describe the trade-off between sensitivity and specificity across different classification thresholds. They are also used to construct the Receiver Operating Characteristic (ROC) curve, from which the Area Under the Curve (AUC) metric is derived, a higher AUC value indicates better discriminative performance of the model.

Results

Table 1 presents the baseline characteristics of participants according to stroke status. Significant differences were observed between stroke and non-stroke groups in most variables. Participants who had experienced a stroke were notably older and had higher average glucose levels and BMI values compared to those without stroke (all $p < 0.001$). A higher prevalence of hypertension and heart disease was also observed among stroke patients. Moreover, individuals with a history of stroke were more likely to be married, self- employed, or engaged in private-sector work, while the distribution of gender, residence type, and smoking status showed smaller differences.

Table 1. Descriptive Analysis

Variable	Non-stroke (mean(sd))/ %	Stroke (mean(sd))/ %	p-value
Age	41.76 ± 22.27	67.71 ± 12.40	<0.001***
Glucose Level	104.00 ± 43.00	134.57 ± 62.46	<0.001***
BMI	28.82 ± 7.91	30.47 ± 6.33	<0.001***
Gender			
Female	58.1%	57.4%	0.870
Male	40.9%	42.6%	
Hypertension			
No	91.7%	71.3%	<0.001***
Yes	8.3%	28.7%	
Heart Disease			
No	95.7%	80.9%	<0.001***
Yes	4.3%	19.1%	
Marriage			
No	35.8%	11.0%	<0.001***
Yes	64.2%	89.0%	
Work Type			
Government job	12.8%	13.4%	<0.001***
Private	57.1%	60.8%	
Self-employed	15.4%	25.4%	
Never worked	0.5%	0%	
Children	14.3%	0.5%	
Residence Type			
Rural	49.3%	47.8%	0.725
Urban	50.7%	52.2%	
Smoking Status			
Formerly smoked	16.6%	27.3%	<0.001***
Never smoked	37.6%	40.2%	
Smokes	14.9%	18.7%	
Unknown	30.9%	13.9%	

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

When evaluating the model performance, we prioritize the recall rate, which is essential in stroke prediction to minimize the false negative rate. According to the results in Table 2, we focused on the predictive performances for the storke class. Among the three models, the Random Forest classifier achieved the highest overall performance, with the highest recall rate (98%) and an AUC of approximately 0.98, indicating excellent predictive capability of stroke and strong overall discriminative ability. The k-Nearest Neighbor (kNN) model also showed excellent predictive performance (Recall= 96%), and a comparable AUC (0.96), making it an efficient yet powerful non-parametric alternative. In contrast, the Logistic Regression model demonstrated lower predictive ability (Recall = 83%) and AUC (0.86), suggesting that its linear decision boundary is less efficient at identifying the true stroke cases in data characterized by nonlinear and interacting predictors.

Table 2: Evaluation Metrics

Model	Accuracy	Precision	Recall	F1-score	AUC (ROC)
Random Forest	0.94	0.93	0.98	0.94	0.98
k-Nearest Neighbors	0.93	0.88	0.96	0.92	0.96
Logistic Regression	0.79	0.77	0.83	0.79	0.86

Table 3 shows the feature importance analysis from the optimized Random Forest model indicates that age is the most dominant predictor of stroke, contributing nearly 35% of the total importance. The next most influential variables are average glucose level and BMI, which reflect metabolic health, suggesting that elevated blood glucose and higher body mass index are critical physiological indicators associated with stroke occurrence. Sociodemographic variables such as marital status (not married), work type (self-employed or government job), and residential area (rural) also show meaningful contributions, capturing lifestyle and environmental effects. Overall, these findings emphasize that both biological (age, glucose, BMI) and behavioral/lifestyle factors jointly influence stroke risk.

Table 3. Feature Importance Analysis

Rank	Feature	Importance
1	Age	0.348
2	Average Glucose Level	0.163
3	BMI	0.138

4	Ever Married = No	0.066
5	Hypertension	0.046
6	Residence Type = Rural	0.034
7	Gender = Female	0.034
8	Work Type = Self-employed	0.030
9	Smoking Status = Never Smoked	0.027
10	Work Type = Govt Job	0.026

Conclusion

This study examined how machine learning can be applied to predict stroke risk by analyzing clinical and behavioral factors, while also exploring its implications for healthcare management and business analytics. Among the models tested, the Random Forest classifier achieved the best performance, with an accuracy of 94% and an AUC of 0.98, demonstrating strong predictive power in identifying individuals at high risk. Feature importance analysis indicated that age, average glucose level, and BMI were the most influential predictors, followed by marital status, hypertension, and work type. These findings suggest that both physiological and lifestyle-related factors contribute meaningfully to stroke prediction, aligning with previous research (Dubow et al., 2011; Rexrode et al., 2022).

When doing the prediction, the random forest performed the best out of all three models. The performance is likely related to the algorithm's ensemble structure. By aggregating the predictions from many decorrelated decision trees built on bootstrap samples and the random subsets of predictors, random forest can approximate complex non-linear and high-order interactions without requiring a prespecified functional form (Breiman, 2001). Moreover, because each tree uses threshold-based splits on the predictor values, the model depends mainly on the ordering rather than the exact magnitude of the observations, which makes it less sensitive to extreme values. These properties are valuable when modelling the heterogeneous clinical data, where the relationships between risk factors and stroke are unlikely to be linear and measurement error and outliers are common.

From a business perspective, integrating predictive models into healthcare operations offers substantial economic and strategic value. Early identification of high-risk individuals enables hospitals, insurance providers, and digital health companies to implement preventive interventions, optimize resource allocation, and reduce treatment costs. Predictive analytics thus provide a foundation for data-driven decision-making and the development of personalized healthcare services.

Despite those promising results, some challenges remain regarding data interpretability, standardization, and privacy protection. First, machine learning models often suffer from limited interpretability, making it difficult for clinicians to understand how some features contribute to an individual patient's risk, which will hinder the clinical implementation and accountability. Second, several ethical and operational challenges have to be considered. In our analysis, the data originates from different hospitals, making the measurement might differ across the data. Representativeness is another concern, as models trained on a dataset that mainly consist older adults, it could be biased when applied to younger populations.

Third, privacy also is an important challenge. When using sensitive healthcare information for model development, it requires compliance with data protection regulations and secure data storage. Future research should focus on developing explainable and scalable ML frameworks and incorporating broader datasets that include diverse populations, and more behavioral and socioeconomic dimensions. Overall, this study demonstrates that leveraging machine learning for stroke prediction holds both clinical benefits and business potential, advancing efficiency and innovation in the healthcare industry.

References

1. Ahammad, T. (2022). Risk factor identification for stroke prognosis using machine-learning algorithms. *Jordanian Journal of Computers and Information Technology*, *8*(3).
2. Ali, A. A. (2019). Stroke prediction using distributed machine learning based on apache spark. *Stroke*, *28*(15), 89–97.
3. Alruily, M., El-Ghany, S. A., Mostafa, A. M., Ezz, M., & El-Aziz, A. A. (2023). A-tuning ensemble machine learning technique for cerebral stroke prediction. *Applied Sciences*, *13*(8), 5047.
4. Asplund, K., Karvanen, J., Giampaoli, S., Jousilahti, P., Niemelä, M., Broda, G., Cesana, G., Dallongeville, J., Ducimetriere, P., & Evans, A. (2009). Relative risks for stroke by age, sex, and population based on follow-up of 18 European populations in the MORGAM Project. Stroke, 40(7), 2319–2326.
5. Banerjee, C., Moon, Y. P., Paik, M. C., Rundek, T., Mora-McLaughlin, C., Vieira, J. R., Sacco, R. L., & Elkind, M. S. (2012). Duration of diabetes and risk of ischemic stroke: The Northern Manhattan Study. Stroke, 43(5), 1212–1217.
6. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
7. Chen, Z. (2023). Stroke risk prediction based on machine learning algorithms. Highlights Sci. Eng. Technol, 38, 932–941.
8. Cruz-Flores, S., Rabinstein, A., Biller, J., Elkind, M. S., Griffith, P., Gorelick, P. B., Howard, G., Leira, E. C., Morgenstern, L. B., & Ovbiagele, B. (2011). Racial-ethnic disparities in stroke care: The American experience: A statement for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke*, *42*(7), 2091– 2116.
9. Donnell, M. J., Xavier, D., Liu, L., Zhang, H., Chin, S., Rao-Melacini, P., Rangarajan, S., Islam, S., Ardila, S., & Foscal, L. (2010). Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): A case-control study. Lancet, 376(9735), 112–123.
10. Dubow, J., & Fink, M. E. (2011). Impact of hypertension on stroke. Current atherosclerosis reports, 13(4), 298-305.
11. Geethanjali, T., Divyashree, M., Monisha, S., & Sahana, M. (2021). Stroke prediction using machine learning. Journal of Emerging Technologies and Innovative Research, 9(6), 710–717.
12. George, M. G., Tong, X., Kuklina, E. V., & Labarthe, D. R. (2011). Trends in stroke hospitalizations and associated risk factors among children and young adults, 1995–2008. Annals of Neurology, 70(5), 713–721.
13. Giles, W. H., Kittner, S. J., Hebel, J. R., Losonczy, K. G., & Sherwin, R. W. (1995). Determinants of black-white differences in the risk of cerebral infarction: The National Health and Nutrition Examination Survey Epidemiologic Follow-up Study. Archives of Internal Medicine, 155(12), 1319–1324.
14. Gillum, R. F. (1999a). Risk factors for stroke in blacks: A critical review. American

Journal of Epidemiology, 150(12), 1266–1274.
15. Gillum, R. F. (1999b). Stroke mortality in blacks: Disturbing trends. Stroke, 30(8), 1711–1715.
16. Guhdar, M., Melhum, A. I., & Ibrahim, A. L. (2023). Optimizing accuracy of stroke prediction using logistic regression. Journal of Technology and Informatics (JoTI), 4(2), 41–47.
17. Hassan, A., Gulzar Ahmad, S., Ullah Munir, E., Ali Khan, I., & Ramzan, N. (2024). Predictive modelling and identification of key risk factors for stroke using machine learning. Scientific Reports, 14(1), 11498.
18. Horenstein, R. B., Smith, D. E., & Mosca, L. (2002). Cholesterol predicts stroke mortality in the Women's Pooling Project. Stroke, 33(7), 1863–1868.
19. Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process, 5(2), 1.
20. Howard, V. J., Kleindorfer, D. O., Judd, S. E., McClure, L. A., Safford, M. M., Rhodes, J. D., Cushman, M., Moy, C. S., Soliman, E. Z., & Kissela, B. M. (2011). Disparities in stroke incidence contributing to disparities in stroke mortality. Annals of Neurology, 69(4), 619–627.
21. Islam, M. M., Akter, S., Rokunojjaman, M., Rony, J. H., Amin, A., & Kar, S. (2021). Stroke prediction analysis using machine learning classifiers and feature technique. International Journal of Electronics and Communications Systems, 1(2), 17–22.
22. Joubert, J., Prentice, L. F., Moulin, T., Liaw, S.-T., Joubert, L. B., Preux, P.-M., Ware, D., Medeiros de Bustos, E., & McLean, A. (2008). Stroke in rural areas and small communities. Stroke, 39(6), 1920–1928.
23. Kapral, M. K., Fang, J., Hill, M. D., Silver, F., Richards, J., Jaigobin, C., & Cheung, A. M. (2005). Sex differences in stroke care and outcomes: Results from the Registry of the Canadian Stroke Network. Stroke, 36(4), 809–814.
24. Kaptoge, S., Di Angelantonio, E., Lowe, G., Pepys, M., Thompson, S., Collins, R., & Danesh, J. (2010). Emerging Risk Factors Collaboration C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: An individual participant meta- analysis. Lancet, 375(9709), 132–140.
25. Khosla, A., Cao, Y., Lin, C. C.-Y., Chiu, H.-K., Hu, J., & Lee, H. (2010). An integrated machine learning approach to stroke prediction. 183–192.
26. Kissela, B. M., Khoury, J. C., Alwell, K., Moomaw, C. J., Woo, D., Adeoye, O., Flaherty, M. L., Khatri, P., Ferioli, S., & De Los Rios La Rosa, F. (2012). Age at stroke: Temporal trends in stroke incidence in a large, biracial population. Neurology, 79(17), 1781–1787.
27. Kleindorfer, D., Broderick, J., Khoury, J., Flaherty, M., Woo, D., Alwell, K., Moomaw, C. J., Schneider, A., Miller, R., & Shukla, R. (2006). The unchanging incidence and case- fatality of stroke in the 1990s: A population-based study. Stroke, 37(10), 2473–2478.
28. Kokkotis, C., Giarmatzis, G., Giannakou, E., Moustakidis, S., Tsatalas, T., Tsiptsios, D., Vadikolias, K., & Aggelousis, N. (2022). An explainable machine learning pipeline for stroke prediction on imbalanced data. Diagnostics, 12(10), 2392.
29. Kuo, S.-H., Lee, Y.-T., Li, C.-R., Tseng, C.-J., Chao, W.-N., Wang, P.-H., Wong, R.-H., Chen,
30. C.-C., Chen, S.-C., & Lee, M.-C. (2013). Mortality in Emergency Department Sepsis score as a prognostic indicator in patients with pyogenic liver abscess. The American Journal of Emergency Medicine, 31(6), 916–921.
31. Maldonado, S., López, J., & Vairetti, C. (2019). An alternative SMOTE oversampling strategy for high-dimensional datasets. Applied Soft Computing, 76, 380–389.
32. Mezher, M. A. (2022). Genetic folding (GF) algorithm with minimal kernel operators to predict stroke patients. Applied Artificial Intelligence, 36(1), 2151179.

33. Nijman, S. W., Leeuwenberg, A., Beekers, I., Verkouter, I., Jacobs, J., Bots, M., Asselbergs, F., Moons, K. G., & Debray, T. P. (2022). Missing data is poorly handled and reported in prediction model studies using machine learning: A literature review. Journal of Clinical Epidemiology, 142, 218–229.
34. Paul, D., Gain, G., Orang, S., Das, P., & Chaudhuri, A. K. (2022). Advanced random forest ensemble for stroke prediction. Training, 66, 34.
35. Reeves, M. J., Fonarow, G. C., Zhao, X., Smith, E. E., & Schwamm, L. H. (2009). Quality of care in women with ischemic stroke in the GWTG program. Stroke, 40(4), 1127–1133.
36. Rexrode, K. M., Madsen, T. E., Yu, A. Y., Carcel, C., Lichtman, J. H., & Miller, E. C. (2022). The impact of sex and gender on stroke. Circulation research, 130(4), 512-528.
37. Roger, V., Go, A., Lloyd-Jones, D., Benjamin, E., Berry, J., Borden, W., Bravata, D., Dai, S., Ford, E., & Fox, C. (2012). American Heart Association Statistics Committee and Stroke Statistics Subcommittee Executive summary: Heart disease and stroke statistics—2012 update: A report from the American Heart Association. Circulation, 125(1), 188–197.
38. Sailasya, G., & Kumari, G. L. A. (2021). Analyzing the performance of stroke prediction using ML classification algorithms. *International Journal of Advanced Computer Science and Applications*, *12*(6).
39. Seshadri, S., Beiser, A., Pikula, A., Himali, J. J., Kelly-Hayes, M., Debette, S., DeStefano, A. L., Romero, J. R., Kase, C. S., & Wolf, P. A. (2010). Parental occurrence of stroke and risk of stroke in their children: The Framingham study. *Circulation*, *121*(11), 1304– 1312.
40. Shanthi, D., Sahoo, G., & Saravanan, N. (2009). Designing an artificial neural network model for the prediction of thrombo-embolic stroke. International Journals of Biometric and Bioinformatics (IJBB), 3(1), 10–18.
41. Sirsat, M. S., Fermé, E., & Câmara, J. (2020). Machine learning for brain stroke: A review.
42. Journal of Stroke and Cerebrovascular Diseases, 29(10), 105162.
43. Sowjanya, A. M., & Mrudula, O. (2023). Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms. Applied Nanoscience, 13(3), 1829–1840.
44. Stroke Prediction Dataset. (2025). [Dataset]. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
45. Sui, X., Lavie, C. J., Hooker, S. P., Lee, D.-C., Colabianchi, N., Lee, C.-D., & Blair, S. N. (2011). A prospective study of fasting plasma glucose and risk of stroke in asymptomatic men. 86(11), 1042–1049.
46. Tazin, T., Alam, M. N., Dola, N. N., Bari, M. S., Bourouis, S., & Monirujjaman Khan, M. (2021). Stroke disease detection and prediction using robust learning approaches. Journal of Healthcare Engineering, 2021(1), 7633381.
47. Van Asch, C. J., Luitse, M. J., Rinkel, G. J., van der Tweel, I., Algra, A., & Klijn, C. J. (2010). Incidence, case fatality, and functional outcome of intracerebral haemorrhage over time, according to age, sex, and ethnic origin: A systematic review and meta-analysis. The Lancet Neurology, 9(2), 167–176.
48. Vasan, R. S., Beiser, A., Seshadri, S., Larson, M. G., Kannel, W. B., D'Agostino, R. B., & Levy,
49. D. (2002). Residual lifetime risk for developing hypertension in middle-aged women and men: The Framingham Heart Study. Jama, 287(8), 1003–1010.
50. Wongvorachan, T., He, S., & Bulut, O. (2023). A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. Information, 14(1), 54.

51. Wu, Y., & Fang, Y. (2020). Stroke prediction with machine learning methods among older Chinese. International Journal of Environmental Research and Public Health, 17(6), 1828.
52. Yiin, G. S., Howard, D. P., Paul, N. L., Li, L., Luengo-Fernandez, R., Bull, L. M., Welch, S. J., Gutnikov, S. A., Mehta, Z., & Rothwell, P. M. (2014). Age-specific incidence, outcome, cost, and projected future burden of atrial fibrillation–related embolic vascular events: A population-based study. Circulation, 130(15), 1236–1244.
53. Zhou, M., Zhu, L., Wang, J., Hang, C., & Shi, J. (2007). The inflammation in the gut after experimental subarachnoid hemorrhage. Journal of Surgical Research, 137(1), 103– 108.