SNS-Bench: Defining, Building, and Assessing Capabilities of Large Language Models in Social Networking Services

Hongcheng Guo [*1] **Yue Wang** [*2] **Shaosheng Cao** [3] **Fei Zhao** [3] **Boyang Wang** [1] **Lei Li** [4] **Liang Chen** [5]
Xinze Lyu [3] **Zhe Xu** [3] **Yao Hu** [3] **Zhoujun Li** [1]

Abstract

With the rapid advancement of Social Networking Services (SNS), the need for intelligent and efficient interaction within diverse platforms has become more crucial. Large Language Models (LLMs) play an important role in SNS as they possess the potential to revolutionize user experience, content generation, and communication dynamics. However, recent studies focus on isolated SNS tasks rather than a comprehensive evaluation. In this paper, we introduce SNS-BENCH, specially constructed for assessing the abilities of large language models from different **S**ocial **N**etworking **S**ervices, with a wide range of SNS-related information. SNS-BENCH encompasses 8 different tasks such as note classification, query content relevance, and highlight words generation in comments. Finally, 6,658 questions of social media text, including subjective questions, single-choice, and multiple-choice questions, are concluded in SNS-BENCH. Further, we evaluate the performance of over 25+ current diverse LLMs on our SNS-BENCH. Models with different sizes exhibit performance variations, yet adhere to the scaling law. Moreover, we hope provide more insights to revolutionize the techniques of social network services with LLMs. https://github.com/HC-Guo/SNS-Bench.

1. Introduction

> *"The true art of socializing is not merely conversing with others, but establishing resonance with them."* **– Carl Jung**

In recent years, the development of Large Language Models (LLMs) has significantly advanced our ability to understand and generate content across various tasks (OpenAI, 2023; DeepSeek-AI, 2024; Touvron et al., 2023; Yang et al., 2024a; Zhao et al., 2023), including text generation, sentiment analysis, dialogue systems, etc. **Social Networking Services (SNS)**, as key platforms for modern information dissemination, have become major venues for human communication, expression of opinions, and emotional transmission. Content on SNS often embodies personal emotions and social-cultural expressions (Yang et al., 2024b; Jin et al., 2024). Therefore, understanding contents requires not only accurate text interpretation but also a deep exploration of the emotions and cultural contexts embedded within.

Despite the success of LLMs across various domains (Xu et al., 2023; Li et al., 2023b), evaluating their performance in SNS remains challenging (Bandura). Recent studies have explored LLMs in specific SNS applications (Jin et al., 2024; Liu et al., 2024; Alhamed et al., 2024a; Qi, 2024; Yang et al., 2024b; Chen et al., 2024), but these efforts focus on isolated tasks rather than providing a comprehensive evaluation of SNS capabilities. Existing benchmarks lack systematic frameworks, standardized metrics, and datasets that capture the diversity of social media interactions, particularly emotional understanding, social interaction, and cultural context. A comprehensive evaluation is crucial for assessing how well LLMs navigate the nuanced and dynamic nature of SNS, ensuring their effectiveness in real-world applications.

To address this gap, we introduce **SNS-BENCH**, a benchmark specifically designed to evaluate the diverse abilities of LLMs on **S**ocial **N**etworking **S**ervices. SNS-BENCH contains 8 SNS-related tasks (e.g. Note Classcification, Sentiment Analysis, and Comment Interaction). Besides, we propose specific metrics for each task. This holistic evaluation enables us to gain deeper insights into the social attributes of LLMs in real-world social contexts, thereby advancing related applications. The data is sourced from a social platform with over 300M users (we have provided valid data license), covering a wide range of domains such as daily life, entertainment, and social issues. It includes various text types, such as notes, posts, comments, and

[1]CCSE, Beihang University [2]Nanjing University [3]Xiaohongshu Inc. [4]The University of Hong Kong [5]Peking University. Correspondence to: Shaosheng Cao <caoshaosheng@xiaohongshu.com>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

conversations.

We conduct experiments on SNS-BENCH with more than 25+ LLMs, with the following main contributions and findings:

- We introduce **SNS-BENCH**, the pioneering benchmark meticulously crafted for social networking service (SNS) environments. Comprising 8 distinct tasks, it spans an extensive array of sub-domains, posing formidable challenges to contemporary LLMs in the arenas of intricate social interactions and note understanding within the SNS landscape.
- The comprehensive evaluation is conducted spanning different sizes and series. Generally, closed-source models outperform open-source ones. Nevertheless, the performance gap between the leading closed-source model and the top open-source model is quite small, only about 1%.
- Models perform more poorly in tasks that involve understanding complex emotions or long notes, compared with more straightforward tasks such as selecting hashtags for notes.

2. Related Work

Recent advances in LLMs have driven interest in their applications within social networking services (SNS). Our work aligns with two key areas: (1) LLMs in social services, including misinformation detection, rumor analysis, and mental health assessment, and (2) the need for systematic benchmarks to evaluate LLM capabilities in SNS.

LLMs in Social Services Recent work has explored LLMs in specific SNS tasks (Zhou et al., 2024; Zeng et al., 2024c; Törnberg, 2024; Yang et al., 2024b; Bandura; Chen et al., 2024). MM-SOC (Jin et al., 2024) evaluates MLLMs' ability to handle social media tasks, which mainly focuses on misinformation detection and hate speech recognition. Liu et al. (2024) propose an LLM-empowered rumor detection approach to teach LLMs to reason over important clues in news and comments. Prior work has demonstrated the effectiveness of LLMs in extracting evidence and analyzing mental health signals from social media content from Reddit posts (Alhamed et al., 2024a). Such analysis research (Qi, 2024) provides empirical evidence for evaluating their capabilities in social networking services. Comprehensive evaluation of LLM capabilities across diverse SNS aspects remains limited. Existing benchmarks lack systematic frameworks, standardized evaluations, and diverse datasets. SNS-BENCH addresses these gaps by open-sourcing a framework to define, evaluate, and assess LLM capabilities across multiple dimensions of social networking services.

Evaluation Benchmarks for Network Tool Using. Many tool-use benchmarks have been proposed, some including network tools, but they face significant limitations. First, these benchmarks struggle to assess complex social media services involving multiple network tools (Huang et al., 2024; Li et al., 2023b; Patil et al., 2023; Tang et al., 2023). Second, many rely heavily on GPT models, resulting in subjective, unstable outcomes and high costs (Qin et al., 2023; Tang et al., 2023). Finally, critical aspects such as multi-user interaction comments on SNS are often overlooked, leading to incomplete evaluations (Zhuang et al., 2023; Li et al., 2023b; Patil et al., 2023).

In contrast, SNS-BENCH addresses these gaps by encompassing diverse SNS-related scenarios and introducing multi-dimensional evaluation metrics. With automated evaluation and metrics like F1, semantic similarity and ANLS, SNS-BENCH provides a more comprehensive and reliable assessment framework.

3. SNS-BENCH

Overview: In Section 3.1, we outline the specific capabilities of LLMs across various SNS-related scenarios, with a particular emphasis on their performance. Section 3.2 details the pipeline used for curating benchmark data. In Section 4, we evaluate 25+ LLMs and provide analysis. Figure 1 presents an overview of 8 tasks in SNS-BENCH.

3.1. Defining

Defining the capabilities of large language models in Social Networking Services (SNS) is crucial. The core skills for SNS (Zeng et al., 2024b; Alhamed et al., 2024b; He et al., 2024) can be summarized into for parts: ***note comprehension, information retrieval, sentiment and intent analysis, and personalized recommendation***. Below, we outline these capabilities and the corresponding designed tasks. The detailed instruction prompts are in Appendix A.2:

• **Note Comprehension.** Models must understand and analyze textual content within SNS, identifying key information. The designed tasks are: **(1) Note-Taxonomy.** Categorizes notes based on content, with single-choice and multiple-choice subsets. **(2) Note-MRC.** Tests reading comprehension by rephrasing queries, evaluating relevance, extracting key information, and providing reasoning, with simple and complex levels.

• **Information Retrieval.** Effective search functionality is essential in SNS. Models should interpret query intent and retrieve relevant information from large datasets, delivering the most accurate results. Tasks: **(3) Note-NER.** Extracts named entities from note content. **(4) Note-QueryCorr.** Matches queries to content, assessing query-based answers and thematic relevance.

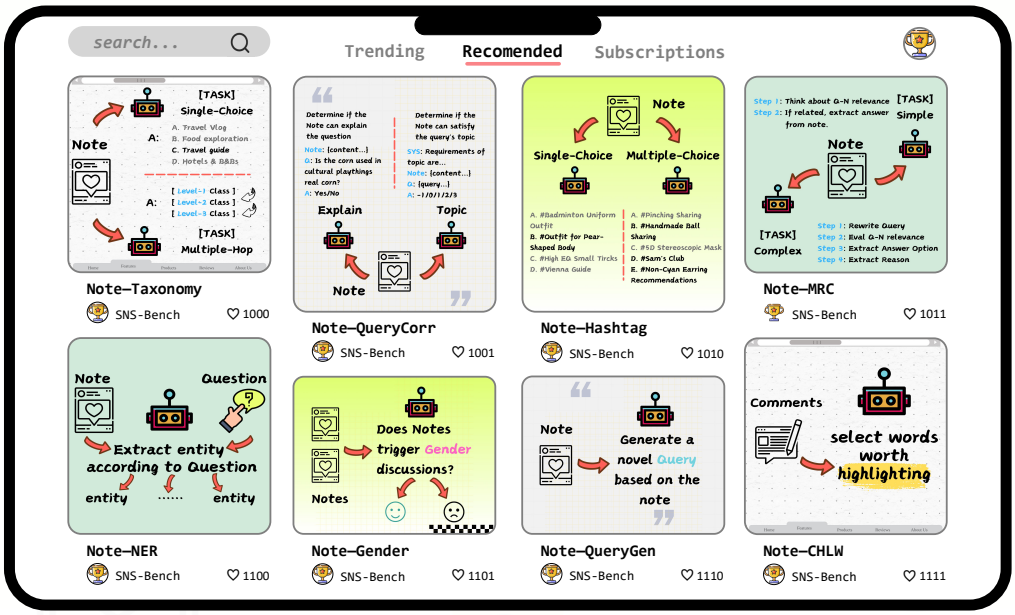

Figure 1: Overview of SNS-BENCH. It contains 8 tasks including Note-Taxonomy, Note-QueryCorr, Note-Hashtag, Note-MRC, Note-NER, Note-Gender, Note-QueryGen, and Note-CHLW.

• **Sentiment and Intent Analysis.** Models should detect emotional tones (e.g., positive, negative, neutral) and underlying intent, especially in sensitive topics like gender discussions. They must identify emotional bias and assess hot-button issues, providing understanding their societal impact. **(5) Note-Gender.** Determines if content is likely to attract attention from both genders, focusing on gender-sensitive topics. **(6) Note-CHLW.** Highlights significant words in the comment section, identifying terms that warrant more attention.

• **Personalized Recommendation.** Models should deliver tailored content recommendations based on user interests, behavior, and past interactions. For instance, the model should suggest relevant content aligned with user preferences. Tasks: **(7) Note-Hashtag.** Selects appropriate hashtags from a given list based on content, in single- and multiple-choice formats. **(8) Note-QueryGen.** Generates more effective search queries based on note content.

3.2. Building

In this section, following the previous benchmark constructions (Chen et al., 2024; Xu et al., 2023; Zhong et al., 2023), we outline the five steps used to build SNS-BENCH and present a detailed distribution analysis.

3.2.1. DATA CONSTRUCTION

The construction of the SNS-BENCH comprises five steps in Figure 2: (1) Data Collection. (2) Processing. (3) Annotation. (4) Quality Control. (5) Expert Review. The quantity distribution of tasks is described in Table 1.

Collection To evaluate the capabilities of large language models in Social Networking Services (SNS), we use data from a major social platform with over 3 billion users [1], hosting diverse user-generated content across various topics. The data undergoes internal company review to ensure compliance and integrity.

We collect user notes, which is the fundamental sharing unit in the platform including the main text, image descriptions, tags, and comments, covering topics such as fashion, travel, food, and lifestyle. During the interaction, other users may leave comments, and the note-owner will categorize the notes with different tags. For evaluating information retrieval and matching, we gather query data and corresponding search results spanning product searches and topic-based queries. User interaction data, including likes, comments, and saves, is compiled to assess personalized recommendation tasks. Additionally, sentiment labels (positive, negative, neutral) and topic tags (e.g., gender discussions, social issues) are included for sentiment and intent analysis, ensuring a comprehensive and representative dataset.

To ensure data diversity and representativeness, we sample across the following standard:

(1) Topic Diversity. Notes are collected from a wide range of categories, including fashion, beauty, health, travel, and

[1]The recent popular REDnote platform.

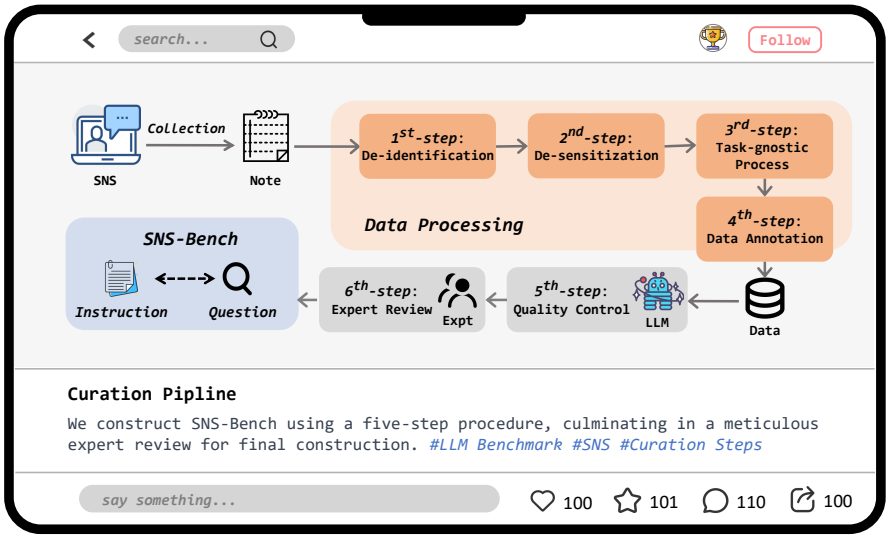

Figure 2: Pipeline for data curation in SNS-BENCH. Mainly five steps: (1) Data Collection. (2) Processing. (3) Annotation. (4) Quality Control. (5) Expert Review.

food, to ensure comprehensive topic coverage.

(2) User Diversity. Data includes users of different ages, genders, locations, and interests, capturing diverse user needs and perspectives.

(3) Time Span. To reflect the dynamic nature of social networks, we focus on real-time content, such as newly published notes and recent comments. Additionally, we include data from 2022–2025 to evaluate the performance over time.

To meet the specific requirements of each task, we collect diverse and targeted data. For **Note-Taxonomy**, notes with varied topic tags are used to evaluate classification capabilities, while **Note-QueryCorr** relies on user search logs with annotated queries to assess query-content matching precision. **Note-Hashtag** utilizes notes with multiple tags to test tag relevance selection, and **Note-MRC** includes notes of varying complexity to measure reading comprehension adaptability. **Note-NER** focuses on entity-rich notes for entity extraction accuracy, and **Note-Gender** are from gender-related discussions. **Note-QueryGen** involves optimizing search terms based on note content, and **Note-CHLW** contains notes with extensive comments to identify key terms effectively. This diverse and carefully curated dataset ensures that we can effectively evaluate the performance of LLMs in a variety of tasks across SNS.

Preprocessing The collected data consists of various formats, primarily text and images. Images are converted to text using Optical Character Recognition (OCR) tools (Li et al., 2023a), if the image is not text-intensive, we will remove it. For challenging cases, manual parsing is employed (Hendrycks et al., 2021; Taylor et al., 2022).

Data undergoes several filtering steps to ensure quality. (1) politically sensitive, inappropriate, or offensive content is removed. (2) user identifiers and personal information are deleted for privacy. (3) irrelevant or low-quality entries, such as those with severe spelling errors, poor grammar, advertisements, or spam, are excluded.

Certain tasks necessitate additional measures. For the Note-Taxonomy task, labels are standardized to eliminate duplicates and ensure consistency. For the Note-Gender task, content with extreme or overly emotional sentiments is filtered out. These preprocessing strategies ensure the data is high-quality and tailored to the needs of specific tasks.

Data Annotation In SNS-BENCH, each data point undergoes a rigorous and systematic annotation process to ensure high-quality labels. 20 annotations are conducted manually to guarantee precision and consistency. For multiple-choice questions, incorrect options are derived from similar but incorrect answers within the notes. Specifically, for the **Note-NER** task, we train a proprietary Named Entity Recognition (NER) model (Wu et al., 2024) to automatically annotate entities, followed by a human review to validate and refine the annotations, ensuring accuracy and reliability. In cases of disagreement, a majority-vote principle is applied. Similarly, for the **Note-CHLW** task, we train a FastText (Bojanowski et al., 2017) classification model to preliminarily identify potential highlight words, which are then manually reviewed to ensure their relevance and accuracy.

More details about our crowdsourcing are in Appendix B.

Quality Control To ensure stringent data quality standards, we adopt a dual-validation approach that combines GPT-4 (OpenAI, 2023) scoring with meticulous manual validation. This process ensures the integrity and reliability of the data while enhancing its overall quality. Using GPT-4 for scoring, we design tailored prompts specific to our dataset, strategically crafted to enable GPT-4 to evaluate and rate the data based on predefined quality criteria. This automated scoring mechanism efficiently identifies and filters out low-quality data instances, while also flagging potential issues and areas for improvement. we adopt FairEval (Wang et al., 2023) to eliminate positional bias. For details on the prompts, please refer to Appendix A.1.

Human Verification Simultaneously, the dataset undergoes rigorous manual validation. A team of expert reviewers (20) conducts an in-depth assessment of each data entry. This process involves cross-validation, where each data point is independently reviewed by at least three different reviewers. Their evaluations focus on content accuracy, coherence, and adherence to domain-specific knowledge. In cases of disagreement, a majority-vote principle is applied, with the final decision reflecting the consensus of the reviewers.

3.2.2. STATISTICS

Table 1: Statistics of SNS-BENCH.

Statistics	Value
Total Questions	6,658
Note-Taxonomy (Single)	1,205
Note-Taxonomy (Multiple)	800
Note-QueryCorr (Explain)	144
Note-QueryCorr (Topic)	800
Note-Hashtag (Single)	800
Note-Hashtag (Multiple)	795
Note-MRC (Complex)	227
Note-MRC (Simple)	800
Note-NER	517
Note-Gender	193
Note-QueryGen	80
Note-CHLW	297
Avg. instruction length (in words)	793
Avg. answer length (in words)	37
Total Input Tokens	4,553,773
Total Output Tokens	1,017,477

Word Cloud (Figure 3). This showcases diverse topics, from consumer themes (***Food, Fashion, Beauty***) to lifestyle interests (***Home, Fitness, Music***). Key terms like ***Topic, Life, Education, Travel*** highlight broad discussions, while interaction-related words (***Comment, Query, Recommendation***) reflect engagement patterns. The mix of practical (***Business, Study***) and leisure (***Art, Travel***) topics ensures comprehensive coverage.

Verb-Noun Pairs (Figure 4). This figure illustrates common verb-noun pairs in social media interactions. Core verbs like ***take, have,*** and ***pay*** form key collocations (***take photo, have time, pay attention***), while peripheral verbs (***solve, share, watch***) add diversity. Nouns related to time (***time, today***), interaction (***message, comment***), and actions (***use, try***) enhance contextual depth.

Length Distribution (Figure 5). Instructions are mostly under ***200*** words, peaking at ***0–25*** words (***1,600*** samples). Few exceed ***1,200***, reflecting detailed cases. Responses are shorter, typically ***10–50*** words, rarely surpassing ***300***, aligning with social media's concise communication style.

4. Experiment

4.1. Evaluation Settings

We conduct experiments on SNS-BENCH using nearly open-source and closed-source LLMs. Uniform prompts are applied across all LLMs in Appendix A.2. Experiments are conducted on 128 NVIDIA H800 GPUs with the OpenCompass codebase (Contributors, 2023).

Evaluated Models Evaluated models include open-source models and closed-source models.

- **Meta Llama 3 Series (Touvron et al., 2023)**: This includes *Llama-3.2-3B*, *Llama-3.2-3B-Instruct*, *Llama-3.1-8B*, *Llama-3.1-8B-Instruct*, and *Llama-3.3-70B-Instruct*.
- **Qwen2.5 Series (Yang et al., 2024a)**: This includes *Qwen-2.5-1.5B*, *Qwen-2.5-1.5B-Instruct*, *Qwen-2.5-7B*, *Qwen-2.5-7B-Instruct*, *Qwen-2.5-32B*, *Qwen-2.5-32B-Instruct*, *Qwen-2.5-72B*, *Qwen2.5-72B-Instruct*
- **Other Open-source Models:** *Phi-4* (Abdin et al., 2024), *Internlm* (Cai et al., 2024), *GLM-4* (Zeng et al., 2024a), *Yi-1.5* (Young et al., 2024) *DeepSeek-V3* (DeepSeek-AI et al., 2024)
- **Closed-source Models:** *GPT4* (OpenAI, 2023), *Gemini* (Anil et al., 2023), *Claude-3.5* (Anthropic, 2024).

4.2. Evaluation Protocol

We choose different metrics for tasks in SNS-BENCH.

For classification tasks including **Note-Taxonomy (Single)**, **Note-QueryCorr (Explain)**, **Note-Gender**, and **Note-Hashtag**. We calculate the accuracy of the final classification. Thus, the formula is:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i). \quad (1)$$

Figure 3: The word cloud of SNS-BENCH.

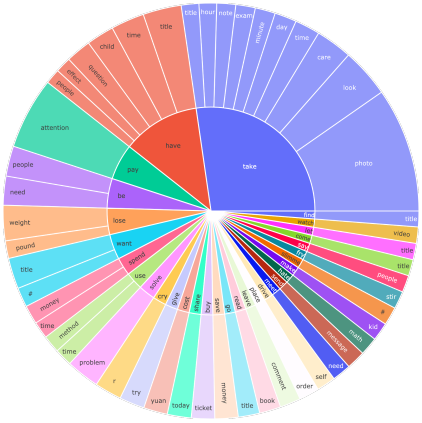

Figure 4: Top 50 Verb-Noun structures in SNS-BENCH instructions.

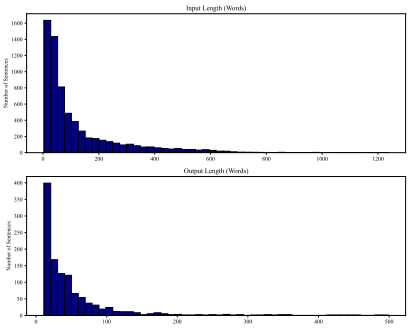

Figure 5: Length distribution of instructions and responses in SNS-BENCH.

where N is the number of samples, $\hat{y}_i$ is the predicted label, y_i is the true label, and $\mathbb{I}$ is an indicator function that is 1 when the predicted label matches the true label, and 0 otherwise.

For the multi-label task **Note-Taxonomy (Multiple)** and **Note-NER**, we compute the F1 score:

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

$$\text{Precision} = \frac{|\hat{y} \cap y|}{|\hat{y}|}, \quad \text{Recall} = \frac{|\hat{y} \cap y|}{|y|}. \tag{3}$$

Here, $\hat{y}$ represents the predicted label set, y denotes the true label set, and $|\cdot|$ indicates the size of a set.

For **Note-QueryCorr (Topic)** task, we calculate the macro-F1 for all categories, which is defined:

$$\text{Macro-F1} = \frac{1}{C} \sum_{i=1}^{C} F1_i \tag{4}$$

where C is the number of categories, and $F1_i$ is the F1 score for category i.

For **Note-MRC (Complex)** task, we first gain the success ratio after parsing inference results. Then, we compute the F1 score and extract match (EM), score between the predicted and true answers. The final score is calculated as:

$$\text{Score} = (\frac{\text{F1} + \text{EM}}{2}) \times \text{SR}, \tag{5}$$

where SR is success ratio. This measures whether the LLMs successfully follow the format instruction.

For **Note-MRC (Simple)**, we first calculate the success ratio SR. Then, we gain the average score of the F1, BLEU and ROUGE scores. The final score is defined as:

$$\text{Score} = \frac{(\text{F1} + \text{BLEU} + \text{ROUGE})}{3} \times \text{SR}, \tag{6}$$

For **Note-QueryGen**, we compute the average of ANLS (Normalized Levenshtein Distance) (Marino et al., 2019) and semantic similarity scores. Semantic similarity is computed using Sentence-BER (Reimers & Gurevych, 2019). The similarity is then determined by the cosine similarity between these embeddings. The overall score formula is as follows:

$$\text{Score} = \frac{\text{ANLS} + \text{SS}}{2}, \tag{7}$$

$$\text{ANLS} = 1 - \frac{D(s, \text{ref})}{\max(|s|, |\text{ref}|)}, \tag{8}$$

$$\text{SS} = \cos(E(s), E(\text{ref})), \tag{9}$$

where SS is Semantic Similarity. $D(s, \text{ref})$ representing the Levenshtein distance between the generated string s and the reference ref, and $|\cdot|$ denoting the length of the string. where $E(\cdot)$ represents the Sentence-BERT embedding function, and $\cos(\cdot, \cdot)$ denotes the cosine similarity between the two embeddings.

For the **Note-CHLW** task, we first calculate the success ratio SR based on the predicted label list. Then, we compute the F1 score between the predicted and true labels. The final score is given by:

$$\text{Score} = \text{F1} \times \text{SR}\ . \tag{10}$$

4.3. Main Results

4.3.1. Overall Evaluation

The benchmark results ***highlight*** key insights into LLM performance on SNS-related tasks.

Closed-Source Models Excel. Proprietary models consistently outperform open-source counterparts, benefit from

superior training resources and optimization techniques. ***Claude-3.5-Sonnet*** leads with an average score of 61.98%, closely follows ***GPT-4o-2024-05-13*** (61.50%).

Task Complexity Varies. Tasks like *Note-Gender* and *Note-Hashtag* show strong results across models, suggesting alignment with current LLM capabilities. In contrast, *Note-MRC (Complex)* and *Note-QueryGen* pose significant challenges, with even top models struggling.

Scale Enhances Performance. Among open-source models, larger architectures like ***Qwen-2.5-72B-Instruct*** achieve the best average score (61.27%), underscoring the importance of model size in tackling diverse SNS tasks, as shown in Table 2.

4.3.2. Task-Specific Evaluation

Our evaluation across eight tasks demonstrates varied model performance patterns. We can conclude: **(1) Imbalance between tasks.** Different models excel in different tasks, highlighting a lack of uniform strength across all SNS-related challenges. Phi-4-14B and Deepseek-V3 achieve strong performance in structured tasks such as Note-QueryCorr, while Claude-3.5-Sonnet leads in tasks requiring contextual understanding like Note-Gender and Note-CHLW. Llama-3.3-70B-Instruct, show noticeable trade-offs, performing exceptionally well in retrieval and taxonomy task. **(2) Task Complexity Variability.** Note-MRC (Complex) and Note-QueryGen remain among the most challenging tasks, as models struggle with deeper comprehension and multi-step reasoning. Note-Taxonomy (Single-Choice) and Note-Hashtag (Single-Choice) exhibit higher scores across most models, indicating these tasks are more approachable for LLMs. **(3) Performance Stability.** Deepseek-V3 and Phi-4-14B demonstrate consistency across related tasks, which maintain balanced performance across Note-Taxonomy, Note-Hashtag, and Note-QueryCorr. GLM-4-9B-Chat shows fluctuating performance, excelling in Note-Taxonomy but underperforming in Note-MRC, suggesting that retrieval-heavy tasks require additional optimization.

5. Analysis

In this section, we first give deeper analysis in three aspects: **Basic Instruction-following, Complex Reasoning (Note-MRC), and Generation (Note-QueryGen).** Then we provide the visualization on Note-QueryGen (Topic) to understand the model preference.

5.1. Error Analysis

We are curious about the root of errors, and examine the basic abilities at a higher level. Thus we dive into basic instruction-following, complex Reasoning (Note-MRC), and generation (Note-QueryGen).

Instruction Following Analysis We analyze instruction following across tasks to assess capability of LLMs in handling structured requirements in diverse social media contexts. Figure 6 shows the success ratio of instruction following across *Note-MRC (Complex)*, *Note-MRC (Simple)*, *Note-CHLW*, and *Note-QueryCorr*. Larger models generally perform better. Task difficulty varies, with complex reading comprehension being the most challenging. Notably, large models still struggle with structured output (**GLM-4**), particularly list generation ([.]), performing significantly below peers and highlighting the impact of strict formatting requirements.

In-depth Analysis in Note-MRC (Complex) To evaluate the complex reasoning capability of large language models on social media content, we analyze their evidence selection and justification performance in the Note-MRC (Complex) task. Instruction is in Appendix A.2. Figure 7 tracks model performance on the complex reading comprehension task. Open-source model ***Deepseek-V3*** leads with an F1 score approaching 40%, followed closely by closed-source models like ***GPT-4o*** and Gemini series.

Bedides, we observe that different models exhibit distinct preferences when initially assessing relevance in Table 3. GPT-series models tend to be more cautious in their judgments, whereas Qwen-series models are comparatively less conservative in determining relevance.

Analysis of Different Metrics in Note-QueryGen Since query generation demands lexical precision and semantic understanding, analyzing these metrics reveals capabilities of models in query comprehension. Figure 8 compares model performance across ANLS, Semantic Similarity, and their average. Semantic Similarity shows little variance between models, limiting its ability to differentiate performance. To address this, we average ANLS and Semantic Similarity, balancing lexical precision and semantic understanding.

5.2. Visualization

In this part, we are curious about the model preference in choosing options.

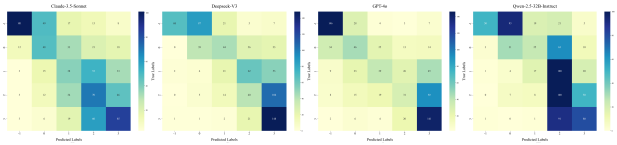

Figure 9: Confusion matrices for Note-QueryGen (Topic) performance of Claude-3.5-Sonnet, Deepseek-V3, GPT-4o, and Qwen-2.5-32B-Instruct.

Figure 9 presents confusion matrices for Note-QueryGen

Table 2: Results of different models on the SNS-BENCH. We utilize green (1st), blue (2nd), and yellow (3rd) backgrounds to distinguish the top three results within both open-source and close-source models.

Models	Note-Taxonomy			Note-Hashtag			Note-QueryCorr			Note-MRC			Note-NER	Note-Gender	Note-CHLW	Note-QueryGen	Avg.
	Single-Choice	Multiple-Hop	Avg.	Single-Choice	Multiple-Choice	Avg.	Explain	Topic	Avg.	Complex	Simple	Avg.	-	-	-	-	-
Open-Source Large Language Models (1.5B+)																	
Qwen-2.5-1.5B	17.51	1.75	9.63	27.50	23.46	25.48	29.86	6.83	18.35	4.74	19.84	12.29	21.63	49.22	27.20	35.81	24.95
Llama-3.2-3B	18.76	9.50	14.13	23.75	38.09	30.92	15.97	3.40	9.69	0.19	25.20	12.70	16.08	52.85	32.97	33.95	25.41
Qwen-2.5-1.5B-Instruct	38.42	5.62	22.02	54.75	32.33	43.54	50.00	19.90	34.95	0.92	22.13	11.53	24.90	49.22	28.89	45.37	32.55
Llama-3.2-3B-Instruct	38.26	9.38	23.82	65.00	58.25	61.63	58.33	11.59	34.96	1.32	29.52	15.42	41.16	83.42	25.32	35.61	40.17
Phi-3.5-Mini-Instruct (3.82B)	46.39	32.75	39.57	63.00	58.03	60.52	61.81	17.47	39.64	12.04	45.38	28.71	27.45	68.91	24.78	42.39	41.50
Open-Source Large Language Models (7B+)																	
Llama-3.1-8B	35.6	28.25	31.93	51.12	65.95	58.54	39.58	3.18	21.38	0.24	20.42	10.33	22.49	54.40	35.99	28.48	32.94
Qwen-2.5-7B	35.52	40.50	38.01	68.25	63.18	65.72	22.92	19.33	21.13	24.95	42.78	33.87	37.11	64.25	32.27	39.67	41.50
Internlm-2.5-7B-Chat	49.88	33.88	41.88	75.38	65.15	70.27	57.64	22.54	40.09	10.95	48.53	29.74	29.99	68.91	24.44	44.35	43.71
Llama-3.1-8B-Instruct	43.73	31.75	37.74	66.38	66.86	66.62	46.53	20.10	33.32	17.79	44.74	31.27	47.10	74.61	26.88	38.60	44.52
Internlm-2.5-20B-Chat	52.95	40.25	46.60	77.88	69.54	73.71	57.64	28.89	43.27	5.25	55.49	30.37	28.43	82.90	26.24	44.09	46.95
Internlm-3-8B-Instruct	52.78	47.38	50.08	79.25	69.60	74.43	57.64	24.06	40.85	26.06	16.27	21.17	38.67	74.09	32.36	44.07	46.96
GLM-4-9B-Chat	51.12	51.38	51.25	76.25	72.44	74.35	59.03	18.91	38.97	24.31	27.44	25.88	43.13	81.87	32.16	45.63	49.15
Qwen-2.5-7B-Instruct	51.12	47.88	49.50	79.75	67.84	73.80	54.17	30.57	42.37	29.34	61.29	45.32	45.41	88.08	33.76	44.65	52.86
Phi-4-14B	54.85	60.38	57.62	80.38	78.74	79.56	56.94	35.70	46.32	37.32	69.45	53.39	44.99	89.12	29.23	44.76	55.62
Open-Source Large Language Models (32B+)																	
Qwen-2.5-32B	33.36	43.88	38.62	78.38	75.18	76.78	43.06	25.13	34.10	24.97	54.02	39.50	41.31	72.54	26.26	42.1	46.40
Qwen-2.5-72B	39.00	35.62	37.31	75.62	61.66	68.64	57.64	29.94	43.79	31.28	51.21	41.25	44.46	75.65	27.53	43.33	47.74
Yi-1.5-34B-Chat	57.10	51.88	54.49	77.38	73.29	75.34	58.33	24.80	41.57	30.30	27.06	28.68	36.56	90.67	22.94	44.70	49.37
Llama-3.3-70B-Instruct	60.75	65.12	62.94	83.62	82.93	83.28	63.89	37.63	50.76	35.40	19.36	27.38	56.09	91.19	33.58	46.41	56.45
Qwen-2.5-32B-Instruct	56.68	63.12	59.90	84.12	76.90	80.51	59.03	32.96	46.00	38.97	71.11	55.04	54.51	90.67	38.84	45.66	58.89
Qwen-2.5-72B-Instruct	60.58	66.75	63.67	86.25	84.60	85.43	63.19	38.63	50.91	38.01	71.92	54.97	54.37	92.23	42.24	46.32	61.27
Deepseek-V3	62.16	72.38	67.27	87.25	85.93	86.59	59.03	36.39	47.71	48.67	73.26	60.97	56.00	90.16	40.45	46.03	61.90
Closed-Source Large Language Models (API)																	
GPT-4o-Mini-2024-07-18	55.44	56.50	55.97	83.12	81.84	82.48	65.28	42.61	53.95	41.34	66.82	54.08	53.22	80.31	32.84	42.84	56.96
GLM-4-Plus	60.41	70.50	65.46	84.50	84.11	84.31	64.58	39.68	52.13	39.20	72.41	55.81	53.16	86.53	30.09	44.68	59.02
Gemini-2.0-Flash-Exp	63.15	72.25	67.70	86.62	86.58	86.60	63.19	34.25	48.72	43.68	69.97	56.83	54.23	89.64	25.15	46.98	59.48
Gemini-1.5-Pro	65.39	73.38	69.39	87.00	80.91	83.96	56.94	39.31	48.13	40.74	71.49	56.12	55.76	88.08	40.99	47.40	61.23
GPT-4o-2024-11-20	62.07	69.5	65.79	85.88	84.07	84.98	59.72	43.85	51.79	45.16	72.61	58.89	54.99	88.08	38.96	47.33	61.35
Claude-3.5-Sonnet	55.68	75.00	65.34	86.88	88.58	87.73	64.58	42.43	53.51	45.92	65.44	55.68	60.44	93.26	33.48	46.38	61.98

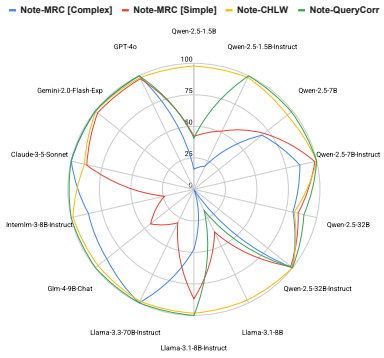

Figure 6: Success ratio (SR) of LLMs in instruction following ability.

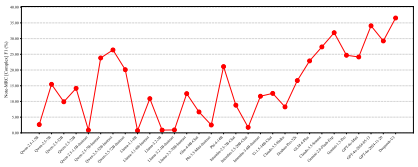

Figure 7: F1 score in Note-MRC (Complex).

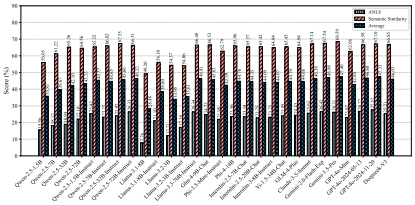

Figure 8: Comparison of ANLS, Semantic Similarity, and average scores on the Note-QueryGen.

Table 3: Model Preferences in Note-MRC (Complex).

Models	Note-MRC (Complex)		
	Precision	Recall	F1
Qwen-2.5-72B-Instruct	34.74	94.29	50.77
GPT-4o-2024-11-20	36.69	88.57	51.88
Deepseek-V3	40.29	80.00	53.59

across four leading models: ***Claude-3.5-Sonnet, Deepseek-V3, GPT-4o,*** and ***Qwen-2.5-32B-Instruct***. The matrices reveal distinct prediction patterns for query-note relevance scores from -1 to 3. High concentration along the diagonal indicates generally accurate predictions, particularly for extreme relevance scores (3 and -1). All models show strong performance in identifying highly relevant matches (score 3). However, models commonly struggle with moderate relevance levels (scores 1 and 2), suggesting challenges in capturing nuanced relationships between queries and content. More visualization results are in Appendix F.

6. Conclusion

We present SNS-BENCH, a comprehensive benchmark for evaluating LLMs in Social Networking Services (SNS). Unlike prior work on isolated tasks, SNS-BENCH offers a structured assessment across eight diverse SNS-related tasks, covering 6,658 questions in multiple formats to capture the complexity of social interactions. We evaluate over 25 LLMs, findings highlight strengths, limitations, and areas for improvement in SNS-related tasks.

Impact Statement

We introduce SNS-BENCH, a benchmark for evaluating language models' social networking capabilities. While AI integration in social platforms offers benefits, it also raises concerns about privacy, authenticity, and social dynamics. SNS-BENCH provides transparent assessment metrics to better understand AI's impact on human communication.

By systematically evaluating social understanding, SNS-BENCH helps identify potential risks and informs safeguards for responsible AI deployment. We take precautions to protect user privacy and remove sensitive information during dataset curation.

Ultimately, our goal is to foster AI development that enhances, rather than undermines, authentic social interactions. SNS-BENCH is designed to improve AI safety and reliability in social networking while preserving user trust and community well-being.

Limitations

Evaluation. While SNS-BENCH offers a comprehensive assessment of social networking tasks, it has certain limitations. The benchmark primarily focuses on text-based interactions, potentially overlooking multimodal aspects. Additionally, our metrics may not fully capture the cultural and contextual nuances that shape social media communication.

Empirical. A key limitation is the dataset's geographic and linguistic scope, which is predominantly English-based and may not fully represent global social networking behaviors. Furthermore, the benchmark focuses on contemporary social media formats, which may not account for emerging trends.

Theoretical. Our evaluation emphasizes practical metrics but lacks theoretical guarantees on their ability to capture the full complexity of human interactions. As social media evolves, some tasks may become less relevant, and the link between benchmark performance and real-world effectiveness requires further study.

References

Abdin, M. I., Aneja, J., Behl, H. S., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R. J., Javaheripi, M., Kauffmann, P., Lee, J. R., Lee, Y. T., Li, Y., Liu, W., Mendes, C. C. T., Nguyen, A., Price, E., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Wang, X., Ward, R., Wu, Y., Yu, D., Zhang, C., and Zhang, Y. Phi-4 technical report. *CoRR*, abs/2412.08905, 2024. doi: 10.48550/ARXIV.2412.08905. URL https://doi.org/10.48550/arXiv.2412.08905.

Alhamed, F., Ive, J., and Specia, L. Using large language models (LLMs) to extract evidence from pre-annotated social media data. In Yates, A., Desmet, B., Prud'hommeaux, E., Zirikly, A., Bedrick, S., MacAvaney, S., Bar, K., Ireland, M., and Ophir, Y. (eds.), *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pp. 232–237, St. Julians, Malta, March 2024a. Association for Computational Linguistics. URL https://aclanthology.org/2024.clpsych-1.22/.

Alhamed, F., Ive, J., and Specia, L. Using large language models (llms) to extract evidence from pre-annotated social media data. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pp. 232–237, 2024b.

Anil, R., Borgeaud, S., Wu, Y., Alayrac, J., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Petrov, S., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T. P., Lazaridou, A., Firat, O., Molloy, J., Isard, M., Barham, P. R., Hennigan, T., Lee, B., Viola, F., Reynolds, M., Xu, Y., Doherty, R., Collins, E., Meyer, C., Rutherford, E., Moreira, E., Ayoub, K., Goel, M., Tucker, G., Piqueras, E., Krikun, M., Barr, I., Savinov, N., Danihelka, I., Roelofs, B., White, A., Andreassen, A., von Glehn, T., Yagati, L., Kazemi, M., Gonzalez, L., Khalman, M., Sygnowski, J., and et al. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805, 2023.

Anthropic. Claude 3.5 sonnet. *Anthropic News*, June 2024. URL https://www.anthropic.com/news/claude-3-5-sonnet.

Bandura, A. *Social learning theory*. Prentice-Hall. URL https://books.google.co.jp/books?id=mjpbjgEACAAJ.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguistics*, 5:135–146, 2017. doi: 10.1162/TACL_A_00051. URL https://doi.org/10.1162/tacl_a_00051.

Cai, Z., Cao, M., Chen, H., Chen, K., Chen, K., Chen, X., Chen, X., Chen, Z., Chen, Z., Chu, P., Dong, X., Duan, H., Fan, Q., Fei, Z., Gao, Y., Ge, J., Gu, C., Gu, Y., Gui, T., Guo, A., Guo, Q., He, C., Hu, Y., Huang, T., Jiang, T., Jiao, P., Jin, Z., Lei, Z., Li, J., Li, J., Li, L., Li, S., Li, W., Li, Y., Liu, H., Liu, J., Hong, J., Liu, K., Liu, K., Liu, X., Lv, C., Lv, H., Lv, K., Ma, L., Ma, R., Ma, Z., Ning, W., Ouyang, L., Qiu, J., Qu, Y., Shang, F., Shao, Y., Song, D., Song, Z., Sui, Z., Sun, P., Sun, Y., Tang, H., Wang, B., Wang, G., Wang, J., Wang, J., Wang, R., Wang, Y., Wang, Z., Wei, X., Weng, Q., Wu, F., Xiong, Y., Zhao, X., and et al. Internlm2 technical report. *CoRR*, abs/2403.17297, 2024. doi: 10.48550/ARXIV.2403.17297. URL https://doi.org/10.48550/arXiv.2403.17297.

Chen, Z., Wu, J., Zhou, J., Wen, B., Bi, G., Jiang, G., Cao, Y., Hu, M., Lai, Y., Xiong, Z., and Huang, M. Tombench: Benchmarking theory of mind in large language models. In Ku, L., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 15959–15983. Association for Computational Linguistics, 2024.

Contributors, O. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass, 2023.

DeepSeek-AI. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model, 2024.

DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Zhang, H., Ding, H., Xin, H., Gao, H., Li, H., Qu, H., Cai, J. L., Liang, J., Guo, J., Ni, J., Li, J., Wang, J., Chen, J., Chen, J., Yuan, J., Qiu, J., Li, J., Song, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Xu, L., Xia, L., Zhao, L., Wang, L., Zhang, L., Li, M., Wang, M., Zhang, M., Zhang, M., Tang, M., Li, M., Tian, N., Huang, P., Wang, P., Zhang, P., Wang, Q., Zhu, Q., Chen, Q., Du, Q., Chen, R. J., Jin, R. L., Ge, R., Zhang, R., Pan, R., Wang, R., Xu, R., Zhang, R., Chen, R., Li, S. S., Lu, S., Zhou, S., Chen, S., Wu, S., Ye, S., Ye, S., Ma, S., Wang, S., Zhou, S., Yu, S., Zhou, S., Pan, S., Wang, T., Yun, T., Pei, T., Sun, T., Xiao, W. L., and Zeng, W. Deepseek-v3 technical report. *CoRR*, abs/2412.19437, 2024. doi: 10.48550/ARXIV.2412.19437. URL https://doi.org/10.48550/arXiv.2412.19437.

He, L., Omranian, S., McRoy, S., and Zheng, K. Using large language models for sentiment analysis of health-related

social media data: empirical evaluation and practical tips. *medRxiv*, pp. 2024–03, 2024.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the MATH dataset. In Vanschoren, J. and Yeung, S. (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021.

Huang, Y., Shi, J., Li, Y., Fan, C., Wu, S., Zhang, Q., Liu, Y., Zhou, P., Wan, Y., Gong, N. Z., and Sun, L. Metatool benchmark for large language models: Deciding whether to use tools and which to use. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=R0c2qtalgG.

Jin, Y., Choi, M., Verma, G., Wang, J., and Kumar, S. MM-SOC: Benchmarking multimodal large language models in social media platforms. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6192–6210, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.370. URL https://aclanthology.org/2024.findings-acl.370/.

Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., and Wei, F. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 13094–13102, 2023a.

Li, M., Song, F., Yu, B., Yu, H., Li, Z., Huang, F., and Li, Y. Api-bank: A benchmark for tool-augmented llms, 2023b.

Liu, Q., Tao, X., Wu, J., Wu, S., and Wang, L. Can large language models detect rumors on social media? *arXiv preprint arXiv:2402.03916*, 2024.

Marino, K., Rastegari, M., Farhadi, A., and Mottaghi, R. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 3195–3204. Computer Vision Foundation / IEEE, 2019.

OpenAI. Gpt-4 technical report. *PREPRINT*, 2023.

Patil, S. G., Zhang, T., Wang, X., and Gonzalez, J. E. Gorilla: Large language model connected with massive apis. *arXiv preprint arXiv: 2305.15334*, 2023.

Qi, J. The impact of large language models on social media communication. In *Proceedings of the 2024 7th International Conference on Software Engineering and Information Management*, ICSIM '24, pp. 165–170, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400709197. doi: 10.1145/3647722.3647749. URL https://doi.org/10.1145/3647722.3647749.

Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., Lin, Y., Cong, X., Tang, X., Qian, B., Zhao, S., Tian, R., Xie, R., Zhou, J., Gerstein, M., Li, D., Liu, Z., and Sun, M. Toolllm: Facilitating large language models to master 16000+ real-world apis, 2023.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 3980–3990. Association for Computational Linguistics, 2019. doi: 10.18653/V1/D19-1410. URL https://doi.org/10.18653/v1/D19-1410.

Tang, Q., Deng, Z., Lin, H., Han, X., Liang, Q., and Sun, L. Toolalpaca: Generalized tool learning for language models with 3000 simulated cases. *arXiv preprint arXiv:2306.05301*, 2023.

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., and Stojnic, R. Galactica: A large language model for science. 2022.

Törnberg, P. Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*, pp. 08944393241286471, 2024.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., and Sui, Z. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023.

Wu, C., Ke, W., Wang, P., Luo, Z., Li, G., and Chen, W. Consistner: Towards instructive ner demonstrations for llms with the consistency of ontology and context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19234–19242, 2024.

Xu, Q., Hong, F., Li, B., Hu, C., Chen, Z., and Zhang, J. On the tool manipulation capability of open-source large language models. *arXiv preprint arXiv: 2305.16504*, 2023.

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024a.

Yang, D., Ziems, C., Held, W., Shaikh, O., Bernstein, M. S., and Mitchell, J. C. Social skill training with large language models. *CoRR*, abs/2404.04204, 2024b.

Young, A., Chen, B., Li, C., Huang, C., Zhang, G., Zhang, G., Li, H., Zhu, J., Chen, J., Chang, J., Yu, K., Liu, P., Liu, Q., Yue, S., Yang, S., Yang, S., Yu, T., Xie, W., Huang, W., Hu, X., Ren, X., Niu, X., Nie, P., Xu, Y., Liu, Y., Wang, Y., Cai, Y., Gu, Z., Liu, Z., and Dai, Z. Yi: Open foundation models by 01.ai. *CoRR*, abs/2403.04652, 2024.

Zeng, A., Xu, B., Wang, B., Zhang, C., Yin, D., Rojas, D., Feng, G., Zhao, H., Lai, H., Yu, H., Wang, H., Sun, J., Zhang, J., Cheng, J., Gui, J., Tang, J., Zhang, J., Li, J., Zhao, L., Wu, L., Zhong, L., Liu, M., Huang, M., Zhang, P., Zheng, Q., Lu, R., Duan, S., Zhang, S., Cao, S., Yang, S., Tam, W. L., Zhao, W., Liu, X., Xia, X., Zhang, X., Gu, X., Lv, X., Liu, X., Liu, X., Yang, X., Song, X., Zhang, X., An, Y., Xu, Y., Niu, Y., Yang, Y., Li, Y., Bai, Y., Dong, Y., Qi, Z., Wang, Z., Yang, Z., Du, Z., Hou, Z., and Wang, Z. Chatglm: A family of large language models from GLM-130B to GLM-4 all tools. *CoRR*, abs/2406.12793, 2024a. doi: 10.48550/ARXIV.2406.12793. URL https://doi.org/10.48550/arXiv.2406.12793.

Zeng, J., Huang, R., Malik, W., Yin, L., Babic, B., Shacham, D., Yan, X., Yang, J., and He, Q. Large language models for social networks: Applications, challenges, and solutions. *arXiv preprint arXiv:2401.02575*, 2024b.

Zeng, J., Huang, R., Malik, W., Yin, L., Babic, B., Shacham, D., Yan, X., Yang, J., and He, Q. Large language models for social networks: Applications, challenges, and solutions. *arXiv preprint arXiv:2401.02575*, 2024c.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J., and rong Wen, J. A survey of large language models. *ARXIV.ORG*, 2023. doi: 10.48550/arXiv.2303.18223.

Zhong, W., Cui, R., Guo, Y., Liang, Y., Lu, S., Wang, Y., Saied, A., Chen, W., and Duan, N. Agieval: A human-centric benchmark for evaluating foundation models, 2023.

Zhou, H., Hu, C., Yuan, Y., Cui, Y., Jin, Y., Chen, C., Wu, H., Yuan, D., Jiang, L., Wu, D., et al. Large language model (llm) for telecommunications: A comprehensive survey on principles, key techniques, and opportunities. *arXiv preprint arXiv:2405.10825*, 2024.

Zhuang, Y., Yu, Y., Wang, K., Sun, H., and Zhang, C. Toolqa: A dataset for llm question answering with external tools. *arXiv preprint arXiv:2306.13304*, 2023.

Appendices

Within this supplementary material, we elaborate on the following aspects:

A. Prompt Template

A.1. Prompt for Quality Control

Prompt Template used for Quality Control

You are now a data grader. You will grade the data I provide according to my requirements, explain the reasons, and then give a piece of higher-quality data based on this piece of data.
Please help me rate the following dialogue data and explain the reasons. Require:
1. Scoring perspective: whether the problem belongs to the field of social network services; whether the description is clear; whether the answer is accurate; whether the language is coherent;
2. Point scale: 5-point scale, 1 point: very poor; 2 points: slightly poor; 3 points: barely qualified; 4 points: usable; 5 points: excellent. Please rate the answer.
3. Format: You can only return a parsable json format data, no other content. For example: "score": 4, "reason": "". Among them, score represents the score for this question, reason represents the reason for the score, and states the advantages and disadvantages of the data.
4. All reasons are written in reason.

A.2. Prompt Templates for Instructions

Prompt Template used for Note-Taxonomy (Single-Choice)

SYSTEM: Given a SNS note's title and content, and a list of candidate categories, please select the most suitable category from the candidate category list.
Here are some examples:

Input :
Note title: Need help, is this a car dash cam
Note content: Bought a DV for 250, came with a battery and charging cable, and it's very light, wondering if it's really a dash cam[crying R][crying R]#DV[topic]#
Candidate Categories :
Car knowledge, Trending, Car lifestyle, Car accessories, Motorcycles, Car shopping, Car modification, New energy & smart, Car culture, Driving test learning, Other automotive
Answer : Car accessories

Input :
Note content: Traveling Henan · Understanding China | Xinxiang South Taihang Tourism Resort: Let's follow the lens to see the sea of clouds in South Taihang #TravelingHenanUnderstandingChina #HomelandHenan #HomelandHenanNewMediaMatrix #May19ChinaTourismDay #XinxiangSouthTaihangTourismResort #SeaOfClouds #XinxiangCultureBroadcastingForeignAffairsAndTourismBureau
Candidate Categories :
Travel VLOG, Shopping, Food exploration, Places to go, Travel guide, Travel records, Hotels & B&Bs, Attraction experience, Travel tips, Living abroad, Indoor leisure
Answer : Travel records

Please output the answer (in the same format as the examples above) DIRECTLY after 'Answer:', WITHOUT any additional content.

HUMAN:
Input:
{content}
Candidate Categories:
{candidates}
Answer:

Prompt Template used for Note-Taxonomy (Multiple-Hop)

SYSTEM: Given a SNS note's title and content, and lists of candidate primary, secondary, and tertiary categories, please select the most suitable categories from each level.
Answer in the format: Primary category | Secondary category | Tertiary category.
Here are some examples:

Input:
Note title: The super disappointing AHC facial cleanser for me
Note content: During National Day trip to Hong Kong, several colleagues asked me to help buy AHC facial cleanser, saying it was especially good. I followed the trend and bought one, used it in the hotel that night 💥. After washing, my eyes were particularly painful, even a tiny bit made me feel like I was going blind 👉😳 so where exactly is it good 😡😡😱😱. Now a full bottle is sitting at home, won't repurchase. Bought for about 70 HKD ~ equivalent to about 60 RMB. The Innisfree green tea cleanser is still better for an affordable cleanser 😣
Candidate Primary Categories:
['Music', 'Beauty', 'Mother&Baby', 'Outdoor', 'Humanities', 'Photography', 'Gaming', 'Art', 'Trends', 'Entertainment', 'Film&TV', 'Kids', 'Career', 'Food']
Candidate Secondary Categories:
['Education Daily', 'Skincare', 'Other Pets', 'Handicraft', 'Resource Sharing', 'Exhibition', 'Personal Care', 'Shoes', 'Mobile Games', 'Relationship Knowledge', 'Meals', 'Places', 'Language Education', 'Makeup']
Candidate Tertiary Categories:
['Cycling Records', 'Playlist Sharing', 'Cleansing', 'Other Shoes', 'Fruit', 'Self-study', 'Instrument Playing', 'JK', 'Beverage Review', 'Tablet', 'Career Development', 'Boutique', 'Other Campus', 'Driving Safety']
Answer: Beauty | Skincare | Cleansing

Input:
Note title: Day8
Note content: Forgot to post yesterday\nMaking up for it today#ChineseBrushCalligraphyCheckIn[topic]##DailyCalligraphyCheckIn[topic]##
Candidate Primary Categories:
['Anime', 'Beauty', 'Gaming', 'Humanities', 'Home&Decoration', 'Health', 'Relationships', 'Social Science', 'Career', 'Art', 'Education', 'Life Records']
Candidate Secondary Categories:
['Science', 'Car Lifestyle', 'Running', 'Other', 'Culture', 'Weight Loss Medicine', 'Parks', 'Fashion', 'Accessories', 'Weight Loss', 'Music Sharing', 'Finger Gaming']
Candidate Tertiary Categories:
['Bags', 'Fruit', 'Leisure Guide', 'Sheet Music Sharing', 'Font Design', 'Calligraphy', 'Weight Loss Tutorial', 'Snack Review', 'Life Science', 'Swimwear', 'Ball Sports', 'Skincare Collection']
Answer: Humanities | Culture | Calligraphy

Please output the answer (in the same format as the examples above) DIRECTLY after 'Answer:', WITHOUT any additional content.

HUMAN:
Input:
{content}
Candidate Primary Categories:
{candidates_primary}
Candidate Secondary Categories:
{candidates_secondary}
Candidate Tertiary Categories:
{candidates_tertiary}
Answer:

Prompt Template used for Note-Hashtag (Single-Choice)

SYSTEM: Give you a SNS Note title and content, along with a list of candidate topic tags, please select the most relevant tag from the list for the note.
Below are some examples:

Input :
Note Title: ❤️Enfj Big Swords Will Love Life BGM Recommendations 🔥
Note Content: Compiled some BGMs for everyone, if you have any recommended songs, please leave them in the comments section, let's exchange and share together 😀
Candidate Topic Tags :
National Peace and Tranquility, Outfit for Pear-Shaped Body, High EQ Small Tricks, Vienna Guide
Answer : High EQ Small Tricks

Input :
Note Title: Must-Have 🪢 at the Sanya Beach
Note Content: On a summer beach, how can you do without this blue tie-dye dress! 💙 \n \nSize XL, perfectly fits plus-size sisters! 👯 \n \nA high-end, laid-back style, loose design, wear it and instantly look whiter and slimmer! 🪢 \n \nThis dress is not only fashionable but also of great quality, a must-have item for beach vacations! 🏝️ \n \nEven the most picky of you will fall in love with it! 💔
Candidate Topic Tags :
National Trend Short Sleeves, Beach Outfit, Miracle Warm Pictures, Badminton Uniform Outfit
Answer : Beach Outfit

Please output the answer (in the same format as the examples above) DIRECTLY after 'Answer:', WITHOUT any additional content.

HUMAN:
Input:
{content}
Candidate Topic Tags:
{candidates}
Answer:

Prompt Template used for Note-Hashtag (Multiple-Choice)

SYSTEM: Given a title and content of a SNS Note, and a list of candidate topic tags, please select all the tags that match the note from the list of topic tags.
Here are some examples:

Input :
Note Title: Deposit Money in Wuhan
Note Content: Squat for a Wuhan customer manager \n Deposit for 5 years!!!
Candidate Topic Tags :
September Birthday, DIY in Guangzhou, Savings Check-in, Wuhan, Henna Tattoo, Nap Time, Vancouver Mercedes, Must-Have Outdoor Gear, Bank, Liu Xiang, Savings, 5D Stereoscopic Mask, Sam's Club, Non-Cyan Earring Recommendations
Answer : Wuhan, Savings, Savings Check-in, Bank

Input :
Note Title: First Time Making Handmade Balls, Afraid of Sealing Issues!!
Note Content: Love handmade balls so much! Never thought making them myself would be so much fun! But the glue for sealing is really hard to handle. Use too little and it might not be sturdy, use too much and it might burn [Cry R][Cry R]. How long should I wait before peeling off the film? 🐷Girls interested can check my page, switched to a new account, will be sharing more textured balls\n52🌏u (5 balls)
Candidate Topic Tags :
CPA Exam Review, After Being a Mom, Recruiting Agents, Manicure Wholesale Collaboration, First Time Making Handmade Balls, Korea Trip, Shanghai Yuanxing Huayu Real Estate Group Co., Ltd., Handmade Ball, Pinching Ball Fun, Popular Mask Recommendations, Resume Coaching, Handmade Ball FX, Handmade Ball Sharing, New Driving Style, Pinching Sharing, Basic Skills Competition for Class Teachers, Handmade Ball Original
Answer : Pinching Sharing, Handmade Ball Sharing, Handmade Ball FX, Handmade Ball Original, First Time Making Handmade Balls, Handmade Ball, Pinching Ball Fun

Please output the answer (in the same format as the examples above) DIRECTLY after 'Answer:', WITHOUT any additional content.

HUMAN:
Input:
{content}
Candidate Topic Tags:
{candidates}
Answer:

Prompt Template used for Note-QueryCorr (Explain)

SYSTEM: Below is a question and a document. Please determine if the given document can answer the given question. If it can, reply with "Yes", otherwise reply with "No". Only reply with "Yes" or "No", no additional explanations are needed.
Here are some examples:

Question :
Is the distorted face in the back camera of an iPhone the real you?
Document :
Revealed | Is the you in the original camera really you? |- Distorted features? Facial asymmetry? Poor skin condition? Skewed face? Why do you see all these when you open the original camera on your phone? And then you get anxious: Am I really that ugly? Actually, it's not the case. The original camera on your phone does indeed make you look worse. Let me explain one by one about appearance anxiety #notcamerafriendly
Answer : No

Question :
Is the corn used in cultural playthings real corn?
Document :
Light Macaroon | Beautiful, beautiful | All real corn grown, nature's colorful gifts currently popular cultural plaything corn (naturally grown) undyed
Answer : Yes

HUMAN:
Question:
{question}
Document:
{document}
Answer:

Prompt Template used for Note-QueryCorr (Topic)

SYSTEM: You are a relevance scoring robot. You will receive a query and a SNS note content, and you will score the relevance between the note and the query.
Before scoring, I will first introduce you to the concepts of requirements and topics, and then provide you with rating levels to choose from.

First, I will introduce you to the concepts of requirements and topics.

Requirements :
```
In search engines, users express their search intent through queries. However, due to factors such as knowledge level and cognition, the input queries may not accurately describe users' true needs. Therefore, search systems need to accurately understand the real needs behind queries to retrieve the content users need.
Generally, queries can be divided into "Precise Need Queries" and "General Need Queries" based on requirements. Their definitions and differences are as follows:
* Precise Need Queries: queries with clear and unique intentions.
e.g. how to treat knee pain, how to cook tomato and eggs
* General Need Queries: queries that may have multiple intentions, which can be further divided into two categories according to primary and secondary needs:
** Primary Needs: The most direct and basic expectations when users search, covering common user needs. Usually what users first think of and most want to get answers to immediately.
** Secondary Needs: Some additional needs around the primary needs, or specific needs. Although these needs aren't common to all users, they are very important to some users.
```

Topics :
```
Can be understood as core issues or involved entity objects, and notes' topics can be viewed from different angles, examples as follows:
Example 1:
618 must-haves!!Sharing happiness-boosting bathroom fixtures 😌
Abstract perspective: Home bathroom essentials
Specific perspective: Introduction of bathroom cabinet, smart toilet, shower head

Example 2:
Five easy-to-raise dogs that don't smell
Abstract perspective: Introduction to easy-to-raise dogs
Specific perspective: Introduction of 5 dogs: Schnauzer, Poodle, Bichon Frise, Shiba Inu, Pomeranian

Based on the matching degree between "query topic" and "note topic", they can be divided into 3 categories:
* Category 1: Complete topic match: The note's topic matches the query topic For example:
Query=home bathroom recommendations, Note=618 must-haves!!Sharing happiness-boosting bathroom fixtures 😌
Query=which dogs are easy to raise, Note=Five easy-to-raise dogs that don't smell

* Category 2: Partial topic match: Part of the note's topic matches the query topic
For example:
Query=smart toilet, Note=618 must-haves!!Sharing happiness-boosting bathroom fixtures 😌
Query=poodle, Note=Five easy-to-raise dogs that don't smell

* Category 3: No topic match: The note's topic doesn't match the query topic at all
For example:
Query=ceiling decoration recommendations, Note=618 must-haves!!Sharing happiness-boosting bathroom fixtures
```

😳
Query=are kangaroos easy to raise, Note=Five easy-to-raise dogs that don't smell
```

*Available Rating Levels* :
```
* 3: Meets primary needs, complete topic match (relevant content ≥ 80%)
Examples:
Query=basketball Note=Who invented basketball?

* 2: Meets primary needs, partial topic match (relevant content between 10%~80%) Examples:
Query=Schnauzer vs Poodle comparison Note=Five easy-to-raise dogs that don't smell | [Meets secondary needs, complete topic match (relevant content ≥ 80%)]
Query=Apple Note=After becoming famous, Fan Bingbing's most regrettable movie "Apple" | [Meets secondary needs, partial topic match (relevant content between 10%~80%)]
Query=Xiaomi su7 Note=Help! Should I choose su7 or Mercedes c260

* 1: Low satisfaction level, relevant content less than 10%
Examples:
Query=23-week glucose tolerance Note=Peking Union Medical College Hospital International Birth Record - Prenatal Care | [Low satisfaction level, only mentioned]
Query=Can gold be purchased Note=Today's gold price-May 18, 2024 <Up> | [Low satisfaction level, extremely specific need]
Query=How to play badminton Note=About sharing a racket with my bestie to play badminton | [Low satisfaction level, helpful for query with matching topic]
Query=Quick weight loss exercise Note=100 reps daily, anytime anywhere💦belly fat + thigh fat reduction ❗ effective

* 0: Doesn't meet needs, has some connection, keywords match
Examples:
Query=Gemini wants to break up Note=How long does it take for Gemini to forget ex [Doesn't meet needs, has some connection, no keyword match]
Query=Chaotianmen Hotpot Note=Haidilao please stop posting on social media

* -1: Doesn't meet needs, no connection at all
Examples:
Query=Amazing girls Note=These three makeup schools are not recommended, ordinary girls can't afford
```

**HUMAN**:
*The Query and note you received are*:
{content}

Please output the rating level, choosing from the five numerical levels: -1, 0, 1, 2, 3.
Output only the numerical rating, without any additional content.

*Answer*:
```

Prompt Template used for Note-MRC (Complex)

SYSTEM: Task: First rewrite the Query, then perform reading comprehension.

Reference Content Format
* The reference content consists of three parts: Query, Doc, and Level
* Query is the user's search term with question-asking intent; Doc is the main text part of the image note, wrapped in <Doc> and </Doc>; Level defines the expected answer type, indicating what type of answer the user needs.

Task Requirements
Step 1: Based on Query and Level, rewrite the Query into specific answer requirements, please output the rewritten rewrite_query. Here are some examples for the first step:
1. [Query=World's most beautiful museums recommendations, Level=museum names] => [Names of the world's most beautiful museums]
2. [Query=Coody Coffee recommendations, Level=drink names] => [Delicious drink names at Coody Coffee]
3. [Query=Qingdao food, Level=dishes] => [Delicious dishes in Qingdao]
4. [Query=Games similar to Number Bomb, Level=game names] => [What games are similar to Number Bomb]
5. [Query=Birthday cake recommendations in Licheng District, Jinan, Level=bakery cake names] => [Delicious birthday cakes from bakeries in Licheng District, Jinan]
6. [Query=Guangzhou Yuexiu District food recommendations, Level=restaurant dishes] => [Delicious restaurant dishes in Yuexiu District, Guangzhou]
7. [Query=Hiking survival bracelet recommendations, Level=brand model] => [Brand and model recommendations for hiking survival bracelets]
8. [Query=Must-buy in Japan, Level=brand model] => [Brand and model of Japanese specialty products]
9. [Query=iPhone 11 photography recommendations, Level=model] => [iPhone 11 models good for photography]
10. [Query=1000 desk essentials, Level=product names] => [1000 useful products for desks]
11. [Query=Changsha special breakfast recommendations, Level=restaurants] => [Names of restaurants with special breakfast in Changsha]
12. [Query=Nanjing surrounding tours, Level=cities] => [Cities suitable for tourism around Nanjing]
13. [Query=Cheap clothing places in Shenzhen, Level=shopping locations] => [Names of shopping locations for cheap clothes in Shenzhen]
14. [Query=Harbin bathhouse recommendations, Level=bathhouse names] => [Names of worthwhile bathhouses in Harbin]

Step 2: Evaluate the relevance between Doc and rewrite_query, paying attention to important qualifiers in rewrite_query, such as time and location modifiers. If Doc misses important qualifiers, it can be considered irrelevant. Analyze the relevance in 1-2 sentences.

Step 3: If Doc is relevant to rewrite_query, extract all answers (Options) from Doc that can answer the user's question and put them in AnswerList. Options must match the answer type required by rewrite_query. Specifically, if rewrite_query requires "brand model", ensure each answer Option includes both brand name and "model"; if this cannot be satisfied, consider it unextractable. If rewrite_query requires "restaurant dishes", ensure each answer Option includes both "restaurant name" and "dish name". Others like "bakery cake names" follow the same pattern. If Option cannot meet the required answer type, consider it unextractable, and AnswerList will be an empty list [].

Step 4: For each answer Option in AnswerList, extract a related description (Reason) from Doc. Format: ["Option": "xxx", "Reason": "xxx", ...]. If Doc is irrelevant or has no answers, return an empty list [].

Notes
1. Answer Options must follow the original text, non-continuous extraction is allowed.
2. Reasons must be extracted from the original text, no modifications allowed.
3. Answer Options should typically be specific entity nouns or names that can be directly searched through search

engines; avoid extracting vague descriptions like "this hidden gem hotel" or "rice noodles from the shop near XXX bus station".
4. Doc may have various article structures, including a common parallel point structure with leading words. When extracting answer Options, avoid including leading words like "Figure 1", "Figure 2", ..., "Shop 1", "Shop 2", etc.
5. Reason length should not exceed 100 english character widths.

Output Format
<rewrite_query>...</rewrite_query>
<relevance_analysis>...</relevance_analysis>
<AnswerList>["...", ...]</AnswerList>
<Result>["Option": "xxx", "Reason": "xxx", ...]</Result>

HUMAN:
Query:
{query}
Level:
{level}

Doc:
<Doc>
{doc}
</Doc>

Output:

Prompt Template used for Note-MRC (Simple)

SYSTEM: You are a reading comprehension master: I will provide you with a user's note article, including a title and content, where the title may be empty `""`. I will also give you a user's question.

Please Think according to the Following Steps :

* Step 1, combine the note's title and content to determine if the user's question can be answered. If there are specific entity words or qualifiers in the question, they must be satisfied to count as having an answer.
* Step 2, if Step 1 determines that the content can answer the user's question, then extract the answer from the original note content and output the extracted original content.

Content Requirements for Answers :

* Extract sentences that can answer the question, cross-sentence extraction is allowed, but must be complete sentences or paragraphs, multiple consecutive sentences must be extracted continuously, and preserve special characters like.
* If the middle part of consecutive sentences is irrelevant to the answer, it can be skipped for cross-sentence extraction; if the answer is long (more than 120 characters), it can be extracted by points.
* Content must come from the original text, no summarization allowed; do not extract content unrelated to answering the user's question.

Format Requirements for Answers :

* Output the answers in list form, where each value in the list is a sentence or paragraph that can answer the question.
* If no answer can be extracted, return an empty list `[]`.
* Do not output any content other than the answer list.

HUMAN:
Here is the note title and content:
{content}

User's question:
{query}

Answer:

Prompt Template used for Note-NER

SYSTEM: Extract entity information from the Passage according to the Question requirements.
Note that the extracted answers should maintain the same case as in the passage. Here are some examples:

Passage : 212 In stock Japanese DIOR 2023 Fall new single color flame gold blush 100/625/343
Question : Please find all 'season' entities
Answer : Fall

Passage : BUTOO tattoo stickers English series colored tattoo stickers waterproof female long-lasting niche high-end tattoo stickers
Question : Please find all 'style' entities
Answer : high-end, niche

Passage : mmm collection slim fit t-shirt women's versatile solid color curved hem short sleeve v-neck fitted hot girl top
Question : Please find all 'style (for clothing and accessories)' entities
Answer : short sleeve, fitted, curved, v-neck

Passage : Early autumn 2022 new sweater plus-size slimming outfit high-end top skirt autumn winter short skirt two-piece set
Question : Please find all 'category' entities
Answer : skirt, sweater, short skirt

Please output the answer (in the same format as the examples above) DIRECTLY after 'Answer:', WITHOUT any additional content.

HUMAN:
Passage: {content}
Question: {question}
Answer:

Prompt Template used for Note-Gender

SYSTEM: According to the following note content fields, determine whether the content is likely to trigger discussions of interest to both men and women. If it includes such content, reply 'Yes,' otherwise reply 'No.' Only reply with 'Yes' or 'No,' without any additional explanation.

These contents mainly include the following aspects:
* Stereotypes about men and women;
* Gender bias phenomena, gender inequality in occupations, sports, interests, etc.;
* Fertility-related content: including views on childbirth, parenting experiences, etc.;
* Marriage-related content: including matchmaking, weddings, post-marriage life, mother-in-law relationships, marital breakdown, prenuptial property, etc.;
* Romance-related content: emotional experiences, arguments, scumbag men/women, cheating, etc.;
* Sex-related content: including sexual violence, sexual harassment, discussions on sexual knowledge, etc.;

And some related social events.
If the text mentions these topics, it is likely to trigger discussions of interest to both men and women.

Here are some examples:
Note Content :
Title: When I deliberately sleep in separate beds from my boyfriend to see his reaction\nCategory: Emotions-Daily Life\nText: In the end it really scared me #Love[Topic]# #Boyfriend[Topic]# #Couple Daily[Topic]# \nasr:When I deliberately sleep in separate beds from my boyfriend...(rest of content)
Answer : Yes

Note Content :
Title:''The Young Lady's Meng'' \nCategory: Humanities-Reading\nText:#Baby Food[Topic]##Accessories Share[Topic]# \nasr:\nocr:...(rest of content)
Answer : No

HUMAN:
Please provide your judgment on the following content.

Note Content:
{content}
Answer:

Prompt Template used for Note-CHLW

SYSTEM: Please select words worth highlighting from comments. These words should be meaningful and specific entities, issues, or descriptions that spark users' search interest.
For life trivia without meaning or search value and content containing numbers, please select carefully. Here are extraction examples:

1) *Input*: Give me a transparent Bluetooth speaker link; *Return*: `['transparent Bluetooth speaker']`
2) *Input*: Is there a tutorial for Huawei Smart Screen; *Return*: `['Huawei Smart Screen tutorial']`
3) *Input*: Is there any specific fitness and weight loss APK?; *Return*: `['fitness and weight loss APK']`
4) *Input*: Noodles have no calories but the sauce does; *Return*: `['noodles have no calories but sauce does']`
5) *Input*: How's Peking duck; *Return*: `['Peking duck']`

HUMAN:
Comment:
{comment}

Possible Highlight Word Options List:
{possible_list}

Please provide a list of words you think are suitable for highlighting. If you think none are appropriate, please output an empty list `[]`.

Answer:

Prompt Template used for Note-QueryGen

SYSTEM: Assume you are an internet user, your task is: based on the input image/text or video note content, generate 1 novel search term related to the core entity words in the note.

Output Format:
The return result is a single line string, recording the generated search term.
Output Requirements:
1. You need to fully understand the input note information, ensure that the given search terms are complete, novel and interesting, with the possibility of being targeted or in-depth exploration, and can bring incremental information to the current note;
2. Ensure that the given search terms are fluent, without expression problems (including incomprehensible, repetitive, typos, word order, etc.);
3. Good search terms can be questions (such as how to xx, what is xx, how to xx etc.), or appropriate compound words (xx tips, xx tutorials, xx outfits, xx recommendations, xx collection, xx selection), where xx comes from the note content.

HUMAN:
Please generate search terms for the following note based on the above instructions.

Input:
{content}
Search term:

B. Crowdsourcing

In conducting our study, we identified several potential risks to participants. Firstly, there is a risk to privacy and confidentiality, as participants are required to share personal information. To mitigate this, all data will be anonymized and stored securely, with access restricted to authorized personnel only. Secondly, there may be psychological risks, such as discomfort or stress during the tasks. To address this, we have included detailed instructions and debriefing sessions to ensure participants feel supported throughout the process. Additionally, participants have the right to withdraw from the study at any time without penalty. Lastly, while there are no significant physical risks associated with our procedures, we will monitor participants for any signs of distress and provide appropriate support. We pay each participant an hourly rate of $10. The primary participants we recruit are college students with master degree (Age ranging 23-28).

We give the performance of models, and the human check rate is around 12% in Table 4.

Table 4: Task Performance Metrics

Task Name	Base Model Accuracy on SNS-BENCH
Note-NER	88.3%
Note-CHLW	87.6%

C. Cases in SNS-BENCH

For each task, we provide an example in the following content.

The Case of Note-Taxonomy (Single-Choice)

SYSTEM: (See Appendix A.2)

HUMAN:
Input :
Note Title: Happiness for Just 9.9, You Deserve the Wonderful Duck Camera at the Imperial Ancestral Temple!!\n Note Content: The Wonderful Duck Camera is so 🐷! It's to the extent that I want to show it to a plastic surgeon to get my face done like this![Face Pulling H] This year's birthday wish is made in advance: to look like the photo![Shy R][lShy Rl]#WonderfulDuckCamera[Topic]# #AIProfilePhoto[Topic]# #QingHuan[Topic]#
Candidate Categories :
Photography Equipment, Portrait Photography, Photography Skills
Answer : Portrait Photography

The Case of Note-Taxonomy (Multiple-Hop)

SYSTEM: (See Appendix A.2)

HUMAN:
Input :
Note Title: Slim and slender arms can solve 80% of dressing problems!\n Note Content: #Slim Arms [Topic] # #Eliminate Bingo Wings [Topic] # #Strongest Slim Arms [Topic] # #Bingo Wings [Topic] # #Slim Arms and Slim Legs [Topic] # #Arm Shaping Tutorial [Topic] #
Candidate Primary Categories :
['Business & Finance', 'Trends', 'Medical & Health', 'Fitness & Weight Loss', 'Social Sciences', 'Emotions', 'Outdoors', 'Sports', 'Urban Travel']
Candidate Secondary Categories :
['Film & Video Montage', 'Weight Loss', 'Educational Daily', 'Exhibition', 'Football', 'Maternity Experience', 'Posture Correction', 'Anime', 'Immersive Activities']
Candidate Tertiary Categories :
['Skiing Equipment & Apparel', 'Animal Science', 'Other Fitness', 'Other', 'Learning Methods', 'Architectural Design', 'Cultural Collectibles', 'Other Music Performances', 'Weight Loss Tutorials']
Answer : Fitness & Weight Loss | Weight Loss | Weight Loss Tutorials

The Case of Note-Hashtag (Single-Choice)

SYSTEM: (See Appendix A.2)

HUMAN:
Input :
Note Title: Completely Crazy 18.9/6 Bottles!! Secretly Placing an Order Behind the Boss's Back🤫 Hurry and Stock Up\n Note Content: Fino, you really!!\nSuch delicious NFC coconut water, and it's even sold so cheaply?!\nNo extra additives! The ingredient list only has coconut water!\n 0 added sugar, 0 fat and low calories\nA box of 200g is less than 39 calories!!\nPerfect for babies on a weight loss diet😋\nEven workers can easily handle it, lightweight and portable, enjoy anytime!\n↙️ Don't miss this opportunity! Rush now!
Candidate Topic Tags :
#Collagen Bomb, Wool Snatching, Kids' Clothing, Sad Puppy
Answer : Wool Snatching

The Case of Note-Hashtag (Multiple-Choice)

SYSTEM: (See Appendix A.2)

HUMAN:
Input :
Note Title: Making Almond Caramel Today 🍬\n Note Content: \n\t\nMade by my sisters, not by me, but I can eat it [Selfie R]
Candidate Topic Tags :
Caramel Almond Rice Boats, Cream Bars Squeeze, Palau, Glory New School Season Flying to New Life, Gray Industry, Caramel, Good for Insomnia, Cell Light, Hand Account, Small Cities Also Have Beautiful Scenery, Officially Announced Female Love, Precise Location Leak Detection
Answer : Caramel Almond Rice Boats, Caramel

The Case of Note-QueryCorr (Explain)

SYSTEM: (See Appendix A.2)

HUMAN:
Question :
What books to buy for primary school teacher qualification certificate exam
Document :
Primary School Teacher Qualification Books | Teacher Recruitment Exam［Topic］#❤ At that time, I bought books from Zhonggong Education for the written exam and books about interviews from Fenbi for the primary school math. Now I don't need them anymore. (About 90% new) ❤ About the books: actually, the content doesn't change much every year. The exam questions are quite stable. The key areas to focus on are multiple-choice questions. You don't need to memorize a lot, just get familiar with the content. I spent over 300 yuan on a Fenbi offline class for the interview, which guaranteed a refund if I didn't pass. I didn't pass the first time, but I got a refund quickly. The second time, I practiced on my own and passed. Remember to sign up for the interview as soon as the written test results are out.
Answer : Yes

The Case of Note-QueryCorr (Topic)

SYSTEM: (See Appendix A.2)

HUMAN:
Class 3
The Query and note you received are:
` ` `\nQuery: How to draw a simple apple sketch\nNote: Cute simple sketches; Simple sketch Draw a red apple using the number 5 #CuteSimpleSketch[Topic]# #SimpleSketch[Topic]# #DrawASimpleSketch[Topic]# #Christmas[Topic]# Write the number five, connect it with curved lines, draw leaves on it. Add details and color it in, and the red apple is done.\n` ` `

Please output the rating level, choosing from the five numerical levels: -1, 0, 1, 2, 3.
Output only the numerical rating, without any additional content.

Answer: 3

Class 2
The Query and note you received are:
` ` `\nQuery: Zhejiang Middle School Entrance Examination\nNote: Zhoushan middle school admission quota for Zhoushan Middle School. Zhoushan 2023 Zhoushan Middle School admission quota Dinghai Second Middle School 63 students Nanhai 49 students Putuo Wuling 42 students Putuo Chengbei 38 students Zhoushan First Junior High School Donggang Middle School. 27 students 25 students Zhoushan Second Junior High School 13 students Greentown Yuhua 12 students 9 students Dinghai Third Middle School 9 students Jintang Middle School Dinghai Sixth Middle School 7 students Baiquan Middle School 3 students\n` ` `

Please output the rating level, choosing from the five numerical levels: -1, 0, 1, 2, 3.
Output only the numerical rating, without any additional content.

Answer: 2

Class 1
The Query and note you received are:
` ` `\nQuery: Apartment Orientation\nNote: Chengdu Post-95s | My Own Small Apartment is Finally for Rent 7-11; Apartment; IKEA; Furniture; Railway South Station Rental; Rent My Own Small Apartment! Brand new furniture! A full-apartment rental near IKEA at Railway South Station, no agency fee, brand new furniture!! Vintage cream style, fully equipped appliances, utilities included, ready to move in ※ House Details: The community is from Zhonghai, 50 sqm one-bedroom unit facing south to the central courtyard with great natural light, open kitchen, independent bathroom, spacious balcony, and an elevator. ※ Transportation: Railway South Station/Shenxianshu, direct access to subway lines 1, 5, and 7, underground parking/curbside parking/shared bikes, convenient with buses and subways. ※ Surrounding Facilities: Right downstairs there's 7-11/Red Flag Chain Store/Wu Dong Feng, walking distance to IKEA furniture/CapitaLand Mall/RT-MART/Hema Fresh/City First Hospital/Huaxi Hospital/Southern Station Park, located near Railway South and Financial City business districts. Super convenient for dining, shopping, entertainment, and medical needs. ※ #ChengduRentals[Topic]# #RailwaySouthStationRentals[Topic]# #FinancialCityRentals[Topic]# #ChengduOneBedroomRentals[Topic]# #IKEA[Topic]# #ZhonghaiCuipingBay[Topic]# #ChengduOneBedroom[Topic]#\n` ` `

Please output the rating level, choosing from the five numerical levels: -1, 0, 1, 2, 3.
Output only the numerical rating, without any additional content.

Answer: 1

Class 0
The Query and note you received are :
` ` `\nQuery: Shabby House Coffee Second Store\nNote: Chengdu Store Exploration | Vintage French Style Cafe niceland Coffee; Cafe; Vintage; Vintage Style; Chengdu Store Exploration; Store Exploration; French Style; Food Store Exploration Store Name: @niceland cafe Nai Island Coffee ※ Address: No.45, Yixue Alley, Jinjiang District, Chengdu ※ ※ Strolling near Chunxi Road with friends ※ Happened to discover this vintage-style cafe ※ The decor is very romantic and vintage French Castle Style ※ The dim lighting adds a unique atmosphere ※ Every corner is perfect for photos ※ ※ Recommended special coffee ※ The names are very unique ※ My friends and I ordered the Cinderella and Hana Hime series ※ Full of romance and sweetness ※ The combination of coffee and refreshing fruity aroma ※ Is blended very subtly ※ ※ Here, you can not only enjoy coffee during the day ※ But also have a drink with friends in the evening ※ It's a perfect place to visit during the hot summer days ※ ※ #WhereToGoOnTheWeekend[topic]# #ChengduStoreExploration[topic]# #ChengduCafe[topic]# #ChengduStoreExplorationSpot[topic]# #ChengduPhotography[topic]# #ChengduDrinking[topic]# #SpecialtyCoffee[topic]# #Coffee[topic]# #ExploreWithMe[topic]# #100StoreExplorationPlan[topic]#\n` ` `

Please output the rating level, choosing from the five numerical levels: -1, 0, 1, 2, 3.
Output only the numerical rating, without any additional content.

Answer : 0

Class -1
The Query and note you received are :
` ` `\nQuery: Hands swollen from gaming\nNote: Help! My arm got red and swollen badly after getting the hepatitis B shot. Hepatitis B; Hepatitis B vaccine; vaccine It's already the third day, and my arm is still very swollen [crying emoji] Then I went to see a doctor, and they pressed my arm and said it's fine [crying emoji]. They said some people have it worse. If it feels itchy in a few days, it means it's healing [crying emoji]. #HepatitisB[topic]# #HepatitisBVaccine[topic]# #Vaccine[topic]# #HepatitisBAntibody[topic]# #HepatitisBBoosterShot[topic]# #HepatitisBTripleTest[topic]# #TripleTestAllNegative[topic]#\n` ` `

Please output the rating level, choosing from the five numerical levels: -1, 0, 1, 2, 3.
Output only the numerical rating, without any additional content.

Answer : -1

The Case of Note-MRC (Complex)

SYSTEM: (See Appendix A.2)

HUMAN:
Query :
Dog Hair Drying Hands-Free Gadget
Level :
Product Name

Doc :
<Doc>
Hair Dryer\nPet Hair Dryer Neck Mount Stand, Pet Grooming Detangling Brush Hands-Free, Groomers no longer need to hold the hair dryer with their neck, super useful! Pet Groomer
</Doc>

Output :
<rewrite_query>Name of the hands-free gadget for drying dog hair</rewrite_query>\n<relevance_analysis>Doc is relevant to rewrite_query because Doc mentions the pet hair dryer neck mount stand that frees up hands.</relevance_analysis>\n<AnswerList>["Pet Hair Dryer Neck Mount Stand"]</AnswerList>\n<Result>[{"Option": "Pet Hair Dryer Neck Mount Stand", "Reason": "Pet Hair Dryer Neck Mount Stand, Pet Grooming Detangling Brush Hands-Free, Groomers no longer need to hold the hair dryer with their neck, super useful!"}]</Result>

The Case of Note-MRC (Simple)

SYSTEM: (See Appendix A.2)

HUMAN:
Here is the note title and content :
2 years and 10 months old, first time riding a bicycle at Decathlon\nInitially a bit unaccustomed, only able to pedal half circles.\nAfter a while, got the hang of it, awesome.\n14 inches, perfect for a 95 cm tall baby at 2 years and 10 months\nWith training wheels, no fear of falling, yyds\n\n#Every Second is Worth Recording[Topic]# #Children's Balance Bike[Topic]# #Children's Bicycle[Topic]# #Boy's Happiness[Topic]# #Decathlon[Topic]# #Cute Kid's Daily Life[Topic]# #Decathlon Children's Bicycle[Topic]#

User's question :
Can a 2-year-old ride a bicycle

Answer : ["Initially a bit unaccustomed, only able to pedal half circles.\nAfter a while, got the hang of it, awesome.", "14 inches, perfect for a 95 cm tall baby at 2 years and 10 months\nWith training wheels, no fear of falling, yyds"]

The Case of Note-NER

SYSTEM: (See Appendix A.2)

HUMAN:
Passage: Please identify all 'color' entities
Question: Korean Lipstick Liquid Matte Velvet Dirty Orange 220 Pumpkin Color Whitening Lipstick 909 116
Answer: Pumpkin Color, Dirty Orange

The Case of Note-Gender

SYSTEM: (See Appendix A.2)

HUMAN:
Please provide your judgment on the following content.

Note Content:
Title: Yang Li's Stand-up Comedy: I Have a Younger Brother (Continued) 🔥🔥🔥\nCategory: Entertainment-Other Entertainment\nBody: 🔥🔥🔥 Yang Li's stand-up comedy is truly heart-wrenching in the latter half. Mom left the pig farm to my brother, and it's hilarious how women help women. \n\n🌟 Yang Liwei, China's first astronaut, once said, 'Our journey into space is a journey of no return.' Similarly, your devotion to your brother is a one-way street, with no turning back. \n\n👨 'Do you have such a lively sister and brother?' she teased.\n\n#StandUpComedy # Topics #Siblings #Comedy
Answer: Yes

The Case of Note-CHLW

SYSTEM: (See Appendix A.2)

HUMAN:
Comment:
Comment: \nWhat does the tag inside mean, you can see it when buying the XO album, didn't understand what it means [cry]\n\n

Possible Highlight Word Options List:
['tag', 'XO album']

Please provide a list of words you think are suitable for highlighting. If you think none are appropriate, please output an empty list [].

Answer: ['XO album']

The Case of Note-QueryGen

SYSTEM: (See Appendix A.2)

HUMAN:
Please generate search terms for the following note based on the above instructions.

Input:
Note Title: Empty, \nMulti-level classification label for the note: Home Furnishing/Home Appliances/Kitchen Appliances, \nMain content: Many people say that under-sink water purifiers are unhygienic, so I installed a direct drinking water machine. Now they say tea bar machines are good to use. Can't I just buy another tea bar machine? #renovationvlog[topic]# #diningcabinet[topic]#, \nCover image OCR information: Did I fall into another pit? His must have been water in my brain when decorating at that time, \nPopular comments: I just feel that the pipe of the under-sink water purifier is so long, and this pipe is not very hygienic. ; Just boil water with an electric kettle. ; I use this thing; Electric kettle [slyR]; What do you think of this combination [slyR].
Search term: Which is better, tea bar machine or under-sink water purifier

D. More Detailed Results

We provide detailed results of Note-MRC (Simple) in Table 5.

E. Hard Case in Note-Gender

The Hard Case in Note-Gender

SYSTEM: (See Appendix A.2)

HUMAN:
Please provide your judgment on the following content.

Note Content:
Title: All Celebrities Are Innocent, Please Let Them Go\nCategory: Entertainment - Celebrity Entertainment News\nBody: #XXX-Celebrity [Topic]# #Celebrity News [Topic]# #Thank You to All Celebrities Above [Topic]# asr: Don't choose me, I'm the one mocked for having a bowl in the fridge by the entire XXX-Show team. Don't choose me, I'm the one mocked for smoking and having peaceful eyes. Don't choose me, I'm the one mocked for being too melodramatic because of my nice voice. Don't choose me, I'm the one called 'water queen' for winning an award. Don't choose me, I'm the one mocked for plastic surgery before and after debut. Don't choose me, I'm the one called 'green tea' without reason. I'm the one who got blacklisted for screaming excitedly after winning a game in XXX-Show.
Answer:

Gold	GPT-4o	Deepseek-V3	Claude	Gemini	GLM-4-Plus	Llama3.3-70B-Instruct	Qwen2.5-72B-Instruct	Phi-4-14B
Yes	No	No	No	No	No	No	No	Yes

F. Confusion Matrix Details

All confusion matrix results are in Figure 10.

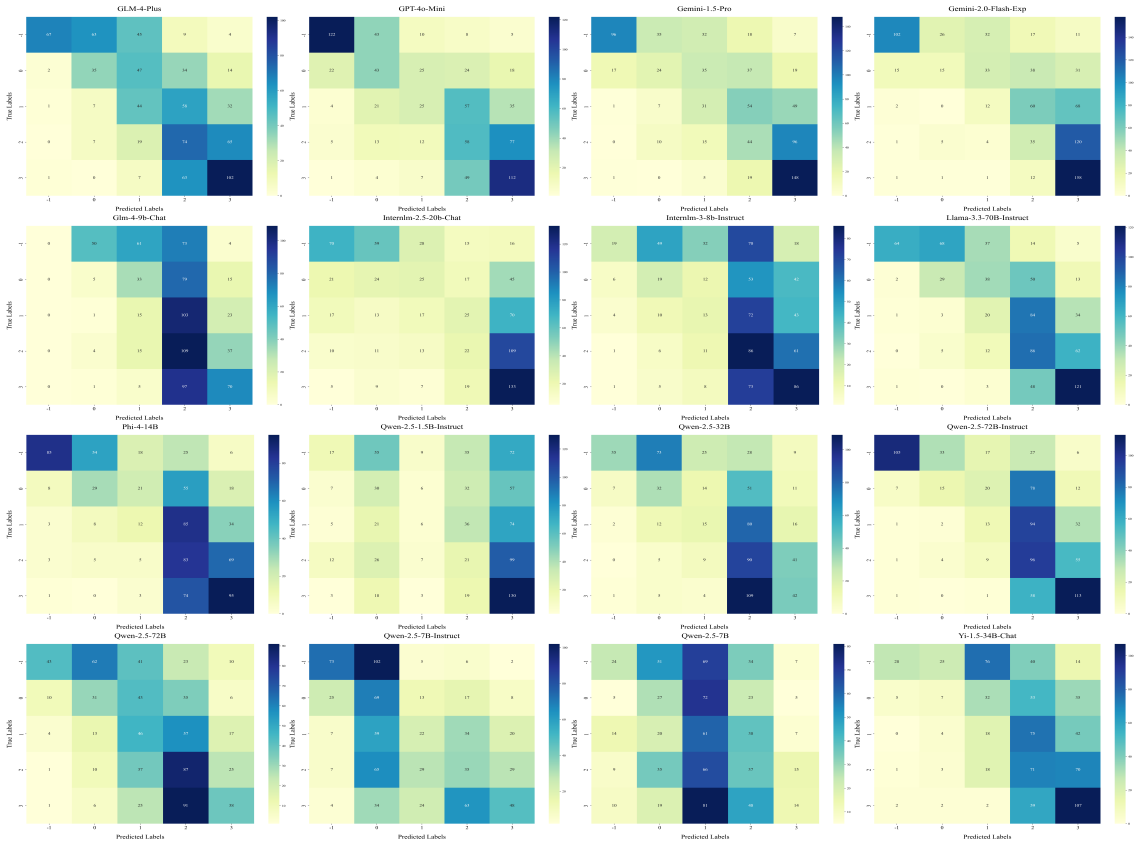

Figure 10: Confusion matrices showing classification performance across different models on Note-QueryCorr task.

Models	Note-MRC (Simple)						
	Success-Ratio	**F1**	**BLEU**	**Rouge-1**	**Rouge-2**	**Rouge-L**	**Overall**
Open-Source Large Language Models (1.5B+)							
Qwen2.5-1.5B	41.87	78.74	30.51	47.59	38.23	41.80	19.84
Qwen2.5-1.5B-Instruct	51.25	73.97	23.42	44.83	34.61	39.04	22.13
Llama-3.2-3B	57.50	76.09	27.46	43.90	33.78	37.89	25.20
Llama-3.2-3B-Instruct	62.30	84.75	29.36	46.19	36.50	40.12	29.52
Phi-3.5-mini-Instruct (3.82B)	72.75	86.42	42.52	65.05	58.07	59.86	45.38
Open-Source Large Language Models (7B+)							
Internlm-3-8B-Instruct	23.88	73.98	55.82	73.49	67.38	70.01	16.27
Llama-3.1-8B	37.75	83.82	34.89	55.04	46.85	49.81	20.42
GLM-4-9B-Chat	43.88	71.53	47.53	67.98	62.18	63.51	27.44
Qwen2.5-7B	71.25	81.56	42.23	62.75	55.65	57.99	42.78
Llama-3.1-8B-Instruct	87.00	9.76	49.62	68.57	63.36	65.85	44.74
Internlm-2.5-7B-Chat	82.12	85.87	40.85	60.96	52.59	55.19	48.53
Internlm-2.5-20B-Chat	92.88	88.84	41.90	60.88	52.03	55.10	55.49
Qwen2.5-7B-Instruct	98.62	84.96	41.67	64.99	58.93	60.17	61.29
Phi-4-14B	93.75	88.08	59.73	77.13	72.32	73.15	69.45
Open-Source Large Language Models (32B+)							
Llama-3.3-70B-Instruct	29.37	36.06	65.39	78.04	74.65	75.44	19.36
Yi-1.5-34B-Chat	38.12	85.18	56.90	73.80	68.98	70.03	27.06
Qwen2.5-32B	84.75	72.33	50.29	68.50	63.10	64.52	54.02
Qwen2.5-72B	76.75	82.87	51.20	69.51	64.35	65.70	51.21
Qwen2.5-32B-Instruct	98.25	75.67	60.90	77.77	73.54	74.00	71.11
Qwen2.5-72B-Instruct	98.25	91.64	57.34	75.08	70.65	71.31	71.92
Closed-Source Large Language Models (API)							
Claude-3.5-sonnet-20241022	87.00	91.98	60.83	77.36	72.79	73.10	65.44
GPT-4o-mini-2024-07-18	99.38	58.92	57.52	75.84	71.41	72.52	66.82
Doubao-Pro-32k	92.62	89.43	61.49	77.96	72.89	74.05	69.62
Gemini-2.0-flash-exp	97.38	90.14	55.03	74.20	69.65	70.28	69.97
GPT-4o-2024-05-13	97.00	84.64	58.23	76.07	71.59	72.51	70.43
Gemini-1.5-pro	97.88	89.69	57.50	75.17	70.94	71.91	71.49
GLM-4-Plus	96.75	90.57	60.13	77.03	72.92	73.54	72.41
GPT-4o-2024-11-20	97.62	88.64	59.99	77.22	72.67	73.37	72.61
Deepseek-V3	96.25	89.63	62.78	78.80	74.55	74.81	73.26

Table 5: Detailed results of Note-MRC (Simple)