SCOP: Evaluating the Comprehension Process of Large Language Models from a Cognitive View

Yongjie Xiao[1, 2], **Hongru Liang**[1, 2*], **Peixin Qin**[1, 2], **Yao Zhang**[3], **Wenqiang Lei**[1, 2]
[1]Sichuan University, China
[2]Engineering Research Center of Machine Learning and Industry Intelligence,
Ministry of Education, China
[3]School of Statistics and Data Science, AAIS, Nankai University, Tianjin, China
xiaoyongjie9@stu.scu.edu.cn lianghongru@scu.edu.cn qinpeixin.scu@gmail.com
yaozhang@nankai.edu.cn wenqianglei@scu.edu.cn

Abstract

Despite the great potential of large language models (LLMs) in machine comprehension, it is still disturbing to fully count on them in real-world scenarios. This is probably because there is no rational explanation for whether the comprehension process of LLMs is aligned with that of experts. In this paper, we propose SCOP to carefully examine how LLMs perform during the comprehension process from a cognitive view. Specifically, it is equipped with a systematical definition of five requisite skills during the comprehension process, a strict framework to construct testing data for these skills, and a detailed analysis of advanced open-sourced and closed-sourced LLMs using the testing data. With SCOP, we find that it is still challenging for LLMs to perform an expert-level comprehension process. Even so, we notice that LLMs share some similarities with experts, e.g., performing better at comprehending local information than global information. Further analysis reveals that LLMs can be somewhat unreliable — they might reach correct answers through flawed comprehension processes. Based on SCOP, we suggest that one direction for improving LLMs is to focus more on the comprehension process, ensuring all comprehension skills are thoroughly developed during training[1].

1 Introduction

Large language models (LLMs) have received much attention in acquiring meanings from documents. Given a document and a question, LLMs are expected to offer the same answer as human experts. However, it is not assured to depend entirely on LLMs in real-world applications (Yang et al., 2024b). One of the possible reasons is that we still don't know whether the comprehension process of LLMs is aligned with experts. As shown in Figure 1(a), an expert answers the question following the process from step 1 to 2 based on the document. LLMs might arrive at the correct answer following other processes, e.g., using memorized data or shortcuts (Figure 1(b)). This difference doesn't matter in a non-existent scenario where LLMs can give correct answers for every question. However, it matters a lot in real-world safety-critical scenarios (law, education, healthcare) especially when normal readers cannot judge the correctness of answers (Pan et al., 2023; Amann et al., 2020). The only solution is to push LLMs to perform the same comprehension process as experts. This encourages us to make a primary attempt and carefully examine the comprehension process of LLMs.

While numerous efforts have been made on comprehension evaluation of LLMs, they are busy finding more proper ways to compare the answers generated after the comprehension process with human references (Rajpurkar et al., 2018; Lai et al., 2017; Dua et al., 2019). This may not provide a reliable judgment, as higher matching scores do not mean better comprehension of the document (Dunietz et al., 2020). A few works (Sugawara et al., 2017; Wang et al., 2022; Dunietz et al., 2020; Sugawara et al., 2021) working in the opposite direction — they prefer to investigate whether LLMs act as accomplished linguists (such as coreference resolution and named entity recognition) before the comprehension process. However, an LLM, good at named entity recognition, may not be good at integrating the whole meaning of a document (Farr and Carey, 1986). Despite various task forms, it remains unclear how LLMs should comprehend the document to close the gap against experts.

To this end, we propose SCOP to carefully study whether LLMs have competitive Skills with experts during the COmprehension Process. We summarize three fundamental challenges behind SCOP: 1) a systematical definition of the requisite skills during the comprehension process, 2) a strict framework

[*]Corresponding author.
[1]https://github.com/SCUNLP/SCOP

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 17407–17431
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

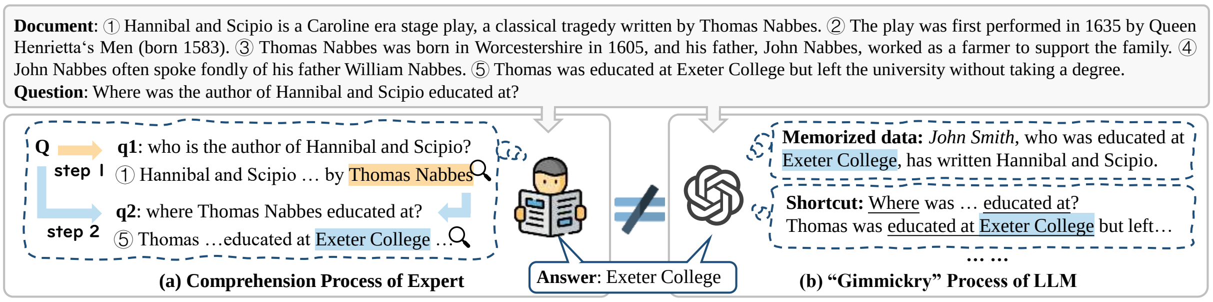
Document: ① Hannibal and Scipio is a Caroline era stage play, a classical tragedy written by Thomas Nabbes. ② The play was first performed in 1635 by Queen Henrietta‘s Men (born 1583). ③ Thomas Nabbes was born in Worcestershire in 1605, and his father, John Nabbes, worked as a farmer to support the family. ④ John Nabbes often spoke fondly of his father William Nabbes. ⑤ Thomas was educated at Exeter College but left the university without taking a degree.
Question: Where was the author of Hannibal and Scipio educated at?
Q
step 1
q1: who is the author of Hannibal and Scipio?
① Hannibal and Scipio … by Thomas Nabbes
step 2
q2: where Thomas Nabbes educated at?
⑤ Thomas …educated at Exeter College …
(a) Comprehension Process of Expert
Answer: Exeter College
Memorized data: John Smith, who was educated at Exeter College, has written Hannibal and Scipio.
Shortcut: Where was … educated at?
Thomas was educated at Exeter College but left…
… …
(b) “Gimmickry” Process of LLM

Figure 1: The comprehension processes of an expert and an LLM. With the same document, question, and answer, the expert gets the answer by first identifying who wrote “Hannibal and Scipio” from ① and then inferring “Exeter College” from ⑤, while LLM might use memorized data, shortcuts, etc.

to construct general testing data for these skills, and 3) a detailed evaluation of LLMs based on the framework. Towards the definition, we decompose the comprehension process into three levels (locating, inferring, interpreting) from local to global based on cognition theories (Krathwohl, 2002; Afflerbach et al., 2015). The interpreting level is further decomposed into three skills (connecting, organizing, selecting). Finally, the comprehension process comprises five skills (cf., Table 1). Towards the framework, we argue the testing data should be independent of formats (e.g., document types, answer styles) and merely focus on the comprehension process. In the absence of suitable data, we introduce a series of strict rules to modify existing datasets and crawl new datasets based on the definitions of skills. The final testing data includes 4,682 samples from 12 datasets. Towards the evaluation, we compare the performances of two close-sourced and two open-sourced LLMs on different levels, skills, document types, and answer styles.

With SCOP, we find that no model has yet achieved the expert-level comprehension process. Besides, aligned with human results (Kintsch and van Dijk, 1978), LLMs are better at local comprehension than global comprehension — better at the locating level than at the inferring and interpreting levels. Surprisingly, our findings with SCOP differ from existing LLM evaluations (Wang et al., 2024; Zheng et al., 2023), where larger closed-sourced models usually outperform open-sourced ones, except for the locating skill. This suggests that, similar to humans, LLMs do not gain better comprehension just by memorizing more information. Further analysis shows that if the comprehension process goes correctly, all LLMs experience a big increase in the inferring skill. One potential improvement for LLMs is to thoroughly learn all comprehension skills during training. We believe evaluating comprehension process provides insightful observations about machine comprehension and gives a new perspective to motivate the future study of LLMs. We hope SCOP can shine a light on how to comprehend documents like experts, accelerating the reliable deployment of LLMs in safety-critical applications. Our contributions are as follows.

- We emphasize the gap between the comprehension process of LLMs and experts, which slows down the real-world application of LLMs.
- We propose SCOP to explore the comprehension process of LLMs from a cognitive view. It includes a systematic definition of five requisite comprehension skills, a strict data construction framework, and a detailed analysis of LLMs.
- With SCOP, we provide insights into how LLMs perform during the comprehension process, facilitating future research in the improvement and deployment of reliable models.

2 Task Definition

The ultimate goal of reading comprehension is to reconstruct information in a document to a meaningful representation in mind and apply it in new situations (Kintsch, 1998). According to Afflerbach et al. (2015), this can be achieved by two processes: the comprehension process, where the reader must obtain local and global meanings of the document; and the thinking process, where the reader must combine these meanings with his background knowledge. In this paper, we focus on the comprehension process, as the gap in background knowledge between LLMs and experts can be narrowed down by feeding more data. From a cognitive view (Krathwohl, 2002; Afflerbach et al., 2015), we decompose the comprehension process into three levels from local to global and propose five skills. Inspired by educational practices (van den Broek and Espin, 2012; Carrell, 1998), we design five tasks specifically tailored

Skill	Task Description	Input	Output
Locating	identify a sentence that supports answering the question	a document and a question	a supporting sentence and answer
Inferring	identify sentences that support answering the question	a document and a question	supporting sentences and answer
Connecting	choose a sentence to connect the previous and following context	a document with blanks and sentence candidates	sentences selected to fill blanks
Organizing	organize the document based on the subheadings	a document with position sequence numbers and subheadings	the positions for subheadings in the document
Selecting	select the key sentences of the document	a document and the desired number of key sentences	key sentences

Table 1: Definitions of the tasks designed for each skill. Data examples are presents in Appendix C.

to evaluate each skill, as explained in Table 1. The detailed definitions are as follows.

Locating The locating level involves questions that should be answered by a piece of information in the document. We treat the sentence containing such information as a supporting sentence. The locating skill focuses on phrase- or sentence-level information and is the basic skill of the comprehension process. To evaluate it, we define a task where the LLM needs to give the supporting sentence and the answer to a question based on the document.

Inferring The inferring level involves questions that should be answered by multiple pieces of information in the document. The inferring skill is more challenging than the locating skill, focusing on sentence- or discourse-level information. To evaluate it, we define a task where LLMs need to identify multiple supporting sentences before answering a question. For example, LLMs should identify ① and ⑤, and then answer with “Exeter College” to the question in Figure 1.

Interpreting The interpreting level involves questions that should be answered by the whole content of a document. This is very similar to making a summary of the document. Instead of directly using the automated summarization task, we care more about the comprehension process before getting the final summary. According to Spivey (1990), this process can be decomposed into the connecting, organizing, and selecting skills:

- With the connecting skill, one can connect discrete messages in a document sentence by sentence. It is a bit like the next sentence prediction task (Devlin et al., 2019), where LLMs should decide whether two sentences appear consecutively in a document. We prefer the sentence cloze task, a more suitable task where LLMs must determine the correct sentence to connect both the previous and the following context.
- With the organizing skill, one can separate the document into several meaningful chunks in logical ways. We design a text segmentation task, where LLMs are required to organize documents based on given subheadings.
- With the selecting skill, one can capture the meaning of a document with several key sentences. We evaluate this skill by tasking LLMs to extract key sentences of the document.

3 Data Construction Framework

Despite numerous efforts to evaluate comprehension, there remains a lack of suitable testing data to evaluate the comprehension process. With the above definitions, we introduce a data construction framework with a series of strict rules to ensure that the testing data reliably evaluates each skill without distractions. Besides, for general testing data, we include both narrative and expository document types, along with various answer styles.

3.1 Locating

We define the locating question as one that focuses on factual details or events explicitly presented in the document. We consider a sentence as the basic unit that encapsulates a fact or an event. Therefore, locating questions can be answered by referencing a single supporting sentence within the document.

We utilize three existing datasets as our source data, each containing questions focused on facts or events: SQuAD v2.0 (Rajpurkar et al., 2018), an expository-type dataset with spans extracted from documents as answers; NewsQA (Trischler et al., 2017), a narrative-type dataset also with span-style answers; and MCTest (Richardson et al., 2013), a narrative-type dataset with multiple-choice answers. Fortunately, SQuAD v2.0 and NewsQA provide annotated spans, making it natural to consider sentences containing these spans as supporting sentences. A direct solution is to treat the questions with only one annotated supporting sentence as locating questions. However, we find this may bring much noise. Some annotated spans do contain the same tokens of answers but are not related to supporting sentences. In Figure 1, sentence ③ contains "Thomas Nabbes" but does not support to answer q_1. Even worse, MCTest only annotates the correct answers, which may not be exactly presented in the document. Thus, the challenge lies in annotating the supporting sentences for questions.

We employ a simple method based on semantic similarity to select the candidate supporting sentences. To find such candidates, the question and its corresponding answer are first transformed into a declarative sentence by prompting Llama3-8B-Instruct (Meta, 2024b). For example, the question "Who's the author of Hannibal and Scipio?" and its answer "Thomas Nabbes" are transformed into "Thomas Nabbes is the author of Hannibal and Scipio." This declarative sentence (d) is then used to retrieve candidate supporting sentences by computing a z-score score with the $i^{th}(0 < i \leq |\mathbb{D}|)$ sentence in the document ($\mathbb{D}$):

$$z_i(d) = \frac{S(d,s_i)-\mu(S(d,s_1),\ldots,S(d,s_{|\mathbb{D}|}))}{\sigma(S(d,s_1),\ldots,S(d,s_{|\mathbb{D}|}))} \qquad (1)$$

where $S(*,\star)$ calculates the cosine similarity between $*$ and $\star$, $\mu(\bullet)$ and $\sigma(\bullet)$ calculates the mean and deviation of $\bullet$, respectively, more details are shown in Appendix A.3. The candidate supporting sentences are ones with z-scores above pre-defined thresholds[2]. We notice this retrieval may be sensitive to answers, particularly answers appearing only once in the document. Thus, we utilize both declarative sentences and questions to retrieve sentences. The final candidate set is the intersection of these retrieval results.

For datasets with annotated spans, if the annotated supporting sentence is in the candidate set, we treat the question as a locating question. For datasets without annotated spans, we use the answer to retrieve a pseudo-supporting sentence. If the retrieved sentence is in the candidate set, we treat the question as a locating question. To verify the pseudo solution, we also apply it on SQuAD v2.0 and NewsQA. We use the questions identified from annotated sentences as gold results and the questions identified from the pseudo solution as predicted results. The F1 scores of the pseudo solution are 0.94 on SQuAD v2.0 and 0.91 on NewsQA.

3.2 Inferring

We define the inferring question as one that should be answered by integrating multiple supporting sentences. We consider three datasets as our source data: HotpotQA (Yang et al., 2024a), an expository-type dataset with span-style answers; MusiQue (Trivedi et al., 2022), also an expository-type dataset with span-style answers; and RACE (Lai et al., 2017), a narrative-type dataset with multi-choice answers. Although each question is annotated with supporting sentences in HotpotQA and MusiQue, not all questions are inferring questions. For example, to answer "Who is the grandpa of Thomas Nabbes?", it requires not only ③ and ④ from the document in Figure 1 but extra commonsense knowledge. To remove such questions, we suppose that an inferring question can be decomposed into more than one subquestion without additional knowledge beyond the document.

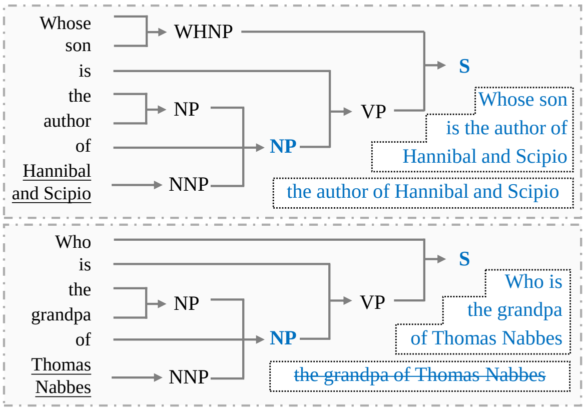

Figure 2: The syntax trees of an inferring question (upper) and non-inferring question (lower).

Thus, for datasets with annotated supporting sentences, we design a semantic-based method to decompose the question into subquestions and identify an inferring question by the number of its subquestions. Specifically, we use the Berkeley Neural Parser[3] to divide the questions into syntax trees. These trees represent how a linear string of words in a question connects to its meaning (Caplan and Hildebrandt, 1988) and help identify whether the question consists of multiple subquestions. We obtain the subquestion candidates by pruning the edges of the "NP" node, at least one of whose child nodes contains a named entity. The question is also included as a root subquestion to prevent any loss of information. In this way, we obtain two candidates for the question in Figure 2 (upper). We note that different candidates may target the same fact and represent the repetitive subquestion, e.g., the two candidates in Figure 2 (lower). To filter out such candidates, we compute the cosine similarity distribution of each candidate over the document

[2] The threshold varies by datasets, we set it to 1 for SQuAD v2.0 and 3 for NewsQA.

[3] https://spacy.io/universe/project/self-attentive-parser/

sentences. We compare the correlation between the distributions of any candidate pairs. If the correlation coefficient is greater than 0.8[4], the candidate with a deeper position in the syntax tree is removed. This method helps us identify inferring questions from bridge questions, where each subquestion contains only a single entity. However, it cannot handle comparison questions, which have subquestion candidates with multiple entities. We identify these questions using their syntactic structure where conjunctions (e.g., "and") connect entities. Some questions (e.g., "Who is older, Queen's Men or Thomas Nabbes) with comparative words like "older" require mathematical knowledge beyond the document. We further utilize part-of-speech tagging "JJR" and "RB" to filter out such questions.

For datasets without annotated supporting sentences (RACE), we prompt GPT-4o-mini (OpenAI, 2024a) to retrieve supporting sentences. Since LLMs might rely on their inherent knowledge, we also perform retrieval based on semantic similarity[5]. We decompose the declarative sentence into sub-sentences, following the same constraints of subquestions. These sub-sentences are then used to retrieve supporting sentences via the solutions proposed in Section 3.1. The final supporting sentences must appear in both the retrieved sets. At last, we identify inferring questions by remaining questions with more than one supporting sentence. To showcase the effectiveness of our method for extracting supporting sentences from narrative-style data, we test on two small-scale datasets with annotated supporting sentences: MultiRC (Khashabi et al., 2018) and OnestopQA (Berzak et al., 2020). The F1 scores for supporting sentence identification are 0.85 and 0.96, respectively, indicating that our method is highly reliable.

3.3 Connecting

We introduce the sentence cloze task to evaluate the connection skill, where LLMs are required to fill in blanks in the document with a set of candidate sentences. We use SCDE (Kong et al., 2020) as our narrative-type dataset and create a new sentence cloze dataset for expository-type data. Specifically, we extract all "Introduction" sections from the ACL OCL Corpus (Rohatgi et al., 2023) as documents. Inspired by Cui et al. (2020), we randomly select five sentences from each document as clozes. These clozes are not consecutive and are not the first or last sentences of a document. We also generate two distractions for each document by randomly sampling candidate sentences from other sections of the paper. After computing the cosine similarity scores with all clozes, we take the top-2 candidates with < 0.6 scores as distractions.

Skill	Source	Size	Answer style	Document type
Locating	SQuAD v2.0	479	Spans	Expository
	NewsQA	984	Spans	Narrative
	MCTest	72	Multi-choice	Narrative
Inferring	HotpotQA	604	Spans	Expository
	MusiQue	510	Spans	Expository
	RACE	547	Multi-choice	Narrative
Connecting	SCDE	625	-	Narrative
	ACL OCL	169	-	Expository
Organizing	ClimateCentral	108	-	Narrative
	WikiHow	146	-	Expository
Selecting	SourceSum	143	-	Narrative
	ACL OCL	295	-	Expository

Table 2: The statistics of our testing data.

3.4 Organizing

The organizing skill is evaluated via a text segmentation task. Instead of just predicting segmentation positions, we require LLMs to organize a document based on given meaningful subheadings. With this constraint, LLMs are more likely to provide consistent segmentation results. While existing text segmentation datasets (Koshorek et al., 2018; Arnold et al., 2019) meet the required format, their subheadings mostly refer to structures (e.g., "Preface") without enough semantic meanings. For narrative-type data, we crawl news documents with subheadings from Climate Central[6]. For expository-type data, we modify the WikiHow (Koupaee and Wang, 2018) dataset by forming related paragraphs as documents and using summaries as subheadings. Each subheading must contain at least four words to serve as a segmentation constraint.

3.5 Selecting

To evaluate the selecting skill, we leverage a key sentence selection task, where LLMs must extract several key sentences from the document to condense its main idea. For narrative-type data, we adopt SourceSum (Suhara and Alikaniotis, 2024), which equips each document with a summary and its source sentences. We treat the source sentences as golden key sentences. For expository-type data, we reuse the "Introduction" data col-

[4] We choose 0.8 as it is commonly used to represent a strong correlation in statistical practice.

[5] A case showing how LLMs rely on their inherent knowledge to retrieve is provided in Appendix B.3.

[6] https://www.climatecentral.org/what-we-do/legal

Model	Locating			Inferring			Connecting		Orgnazing		Selecting	
	MCTest	SQuAD	NewsQA	HotpotQA	MusiQue	RACE	SCDE	ACL	Climate	WikiHow	SourceSum	ACL
Qwen2-72B-Instruct	97.22	98.33	84.15	46.19	11.37	37.29	72.48	33.14	12.96	22.60	16.08	9.83
Llama3.1-70B-Instruct	94.44	99.16	85.67	62.58	30.98	46.07	38.88	11.83	25.93	34.93	16.78	9.83
Claude-3.5-sonnet	97.22	97.08	84.04	56.13	24.71	31.63	69.12	60.95	43.52	55.48	18.18	10.17
GPT-4o	97.22	98.96	89.13	45.36	18.82	37.48	58.72	24.85	27.78	35.62	22.38	12.54

Table 3: Performance of model's comprehension process across five skills.

lected in Sec. 3.3. We further collect the "Abstract" sections for the ACL OCL Corpus (Rohatgi et al., 2023). Each "Abstract" sentence is assigned with the most similar "Introduction" sentence. We treat these assigned sentences as golden key sentences.

3.6 Quality Control and Analysis

The five tasks are designed to ensure LLMs genuinely comprehend the documents when answering, preventing scenarios like "shortcuts". To address cases where LLMs rely on their memory, we further exclude questions that can be directly answered by advanced LLMs (Llama3.1-70B-instruct (Meta, 2024a) and GPT-4o (OpenAI, 2024b)) without referring to the document[7]. Specifically, for datasets with span-style answers, we convert the question into a yes-no question by merging its answer. Then, the LLMs are required to answer this new question. We only keep the question if both answers are "no". For multi-choice datasets, we reuse the strategy proposed in Sec. 3.1 and transform the question with its options into a set of declarative sentences. Then, we require LLMs to select the most sensible one from these declarative sentences. We only keep questions if both selected sentences do not contain the right option. Surprisingly, we filter out more than 90% (33,827) noise samples from 37,023 samples. This further indicates that it is still unsafe to directly deploy LLMs in real-world scenarios, particularly when domain knowledge has conflicts with pre-trained corpora.

During the data construction phase, the API cost is only about \$5, including about \$0.9 for supporting sentence retrieval and about \$4.1 for filtering. Additionally, we randomly select 50 samples from each dataset and ask three workers to check the validity of each sample. A sample gets a score of 1 if all workers agree it is a valid sample; otherwise, the sample is scored 0. The average score achieves 0.81 with an inter-annotator agreement score of 0.73. The statistics of our testing data are shown in Table 2. More details are provided in the Appendix A.

4 Results and Analysis

4.1 Settings

Evaluated Models We involve four LLMs that excel in comprehending documents in SCOP: two open-sourced models (Llama3.1-70B-Instruct (Meta, 2024a) and Qwen2-72B-Instruct (Yang et al., 2024a)) that are deployed directly for inference and two closed-sourced models (GPT-4o (OpenAI, 2024b) and Claude-3.5-sonnet (Anthropic, 2024)) that return answers through API calls[8]. We set the temperature for all LLMs to 0 to ensure controllable results. The prompts for all tasks are listed in the Appendix D.

Metrics We use document-level accuracy (Yessenalina et al., 2010) as the metric for evaluating all skills. The accuracy is defined as the number of times LLM's predictions exactly match the golden results, divided by the total sample size. For the locating and inferring skill, we only focus on the predicted supporting sentences as they reflect the comprehension process of LLMs. The connecting skill requires predicting candidate sentences in the correct sequence. The organizing skill requires predicting the correct positions for subheadings within a document. For the selecting skill, LLMs should predict key sentences.

4.2 Overall Performance

We report the performances of LLMs on five skills in Table 3. It can be seen that all LLMs are far from operating expert-level comprehension processes. Besides, we have the following observations:

Performance w.r.t., different levels LLMs are better at local comprehension than global comprehension. Their performance decreases significantly as the comprehension process moves from local to global levels. Specifically, the average accuracy drops from 93.55% at the locating level to 37.38% at the inferring level, and further to 31.02% at the

[7] This step is deemed necessary, as noted in (Chang et al., 2023) and validated through our pilot experiment in Appendix B.

[8] We also evaluate DeepSeek-R1 (Guo et al., 2025) and GPT-4o-mini, whose performances are consistent with existing results. For LLMs of varying scales, Llama 3.1 7B is tested but excluded from the main evaluation due to its extremely poor performance. More details are in the Appendix B.2.

interpreting level. This suggests while LLMs handle local comprehension well, they struggle with global comprehension. This limitation may result from the next-token prediction task used during pretraining. This training paradigm excels at capturing local contextual information, but it might hinder the model's ability to grasp the bigger picture.

Performance w.r.t., different skills at the interpreting level The three skills at the interpreting level focus on sentences, then discourse, and finally the whole document. The LLMs' performance declines across the three skills, supporting the observation that LLMs are better at local comprehension than global comprehension. For the individual connecting skill, all LLMs perform over twice as poorly on the SCDE dataset compared to the ACL dataset except Claude-3.5. This suggests that most LLMs handle sentence-level logical relationships in narratives better than in expository documents. Besides, LLMs have difficulty with the selecting skill, achieving an average accuracy of only 18.36%. This is likely because the task requires focusing on key sentences spread throughout the document, conflicting with the sequential text processing approach of LLMs[9].

Performance w.r.t., different LLMs Surprisingly, we notice that the performance ranking of different LLMs on SCOP differs from typical LLMs evaluations (Wang et al., 2024; Zheng et al., 2023), where open-sourced models generally outperform closed-sourced ones. Specifically, Llama3.1-70B-Instruct outperforms GPT-4o on the inferring skill, and Qwen2-72B-Instruct exceeds GPT-4o on the connecting skill. This inconsistency may result from the differences between comprehension process evaluation and answer-based evaluation. Comprehension process evaluation examines how LLMs arrive at those answers. This evaluation could reveal bottlenecks in the comprehension process, which helps to find ways to improve and advance LLM's comprehension.

Performance w.r.t., document types Different document types indeed impact LLMs' performance across various skills. For the locating skill, LLMs perform differently on the SQuAD v2.0 and NewsQA datasets, which share the same answer style but differ in document types. LLMs are better at locating facts in expository documents compared to events in narrative documents. In contrast, for connecting and selecting skills, LLMs excel with narrative texts. This is because comprehending the whole content of a narrative is generally easier than expository documents. These results suggest that LLMs' comprehension is somewhat similar to how humans comprehend documents (Graesser, 2003).

Model	Locating	Inferring	Increase
Qwen2-72B-Instruct	4.28	22.50	18.22
Llama3.1-70B-Instruct	4.45	19.68	15.23
Claude-3.5-Sonnet	4.77	12.80	8.03
GPT-4o	2.32	18.29	15.97

Table 4: Inconsistency scores (%) of LLMs on QA data. Scores are averaged across datasets for each skill.

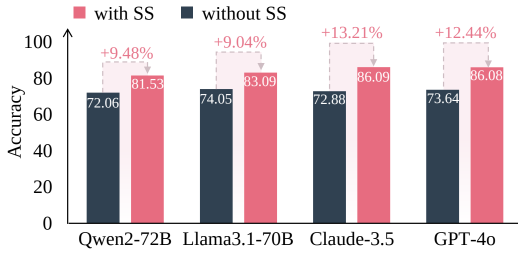

Figure 3: Average accuracy of LLMs with and without supporting sentences (SS) across the inferring datasets.

Performance w.r.t., answer styles The comprehension process of LLMs is influenced by answer styles. When comparing performance on the MCTest and NewsQA datasets, which have the same document types but different answer styles, we observe that LLMs perform better on multiple-choice questions. This might be because the options provided in these questions reduce cognitive load and make comprehension easier. This phenomenon is similar to how humans perform across different answer styles (Frederiksen, 1984).

4.3 Auxiliary Analysis

We further analyze how the comprehension process affects the answer. The results highlight the importance of a correct comprehension process. Besides, we examine the correlation across five skills, validating the rationality of our framework.

Correlation between the comprehension process and answers We analyze the correlation between identifying supporting sentences and answering questions at both the locating and inferring levels. We report the inconsistency score, which is defined as the proportion of samples with incorrect supporting sentences and correct answers. The results in Table 4 reveal that LLMs exhibit inconsistent behaviors even at the basic locating level. This is a strong evidence to support our statements that

[9] A case study of connecting skill is listed in Appendix B.3.

matching answers alone cannot provide a reliable judgment and more efforts are required in comprehension process evaluation. Moreover, we observe that the inconsistency score tends to increase with the complexity of comprehension process. For example, the inconsistency score of GPT-4o is over 7 times higher at the inferring level than at the locating level. After a close look at the failure samples, we find that LLMs may be "slacking off" — they prefer to use unexplainable shortcuts when they cannot identify all the supporting sentences. This emphasizes the importance of studying the comprehension process for more explainable LLMs.

How the comprehension process affects answers A simple question is whether the correctness of comprehension process truely makes a difference for LLMs. This can be investigated at the inferring level, where the golden supporting sentences can construct correct comprehension processes and the predicted answers can used as the observation targets. As shown in Figure 3, with the full set of supporting sentences, the performances of all LLMs significantly increase, suggesting the importance of the comprehension process for LLMs. We further notice that the least lazy LLM (Claude-3.5) gets the biggest benefits (13.21%) from the correct comprehension process. Thus, we believe that it will be much easier to achieve expert-level comprehension by adding more supervision to prevent LLMs from using shortcuts during training.

Correlation among comprehension process skills Finally, we validate our framework by analyzing the correlations in LLMs performance across all datasets, as shown in Figure 4. Our observations are as follows: 1) Datasets evaluating the same skill generally correlate. MCTest is an exception as it is based on simple children's stories. All LLMs perform similarly on this dataset, leading to a lower correlation with others. 2) There is a correlation between locating and inferring skills since inferring builds on locating. The inferring skill involves both locating and aggregating information in two sub-stages. 3) As expected, all three skills at the interpreting level correlate. 4) Some skills are negatively correlated as they emphasize different aspects. For instance, while both connecting and selecting involve comprehending the entire document, connecting focuses on relationships between sentences for a broader view, whereas selecting targets key information for a more centralized view.

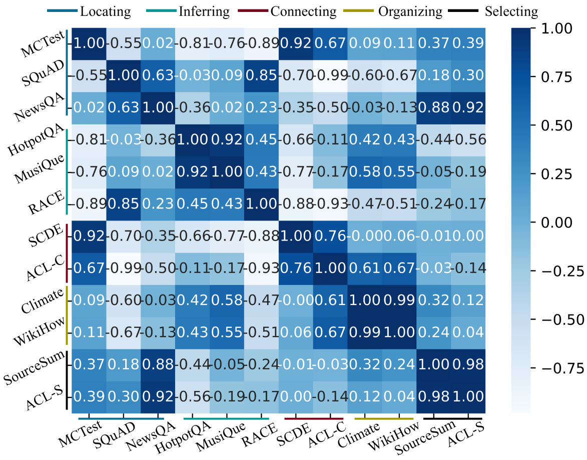

Figure 4: Inter-task performance correlation.

5) Tasks at the interpreting level correlate with both locating and inferring skills, indicating that global comprehension relies on the skills required for local comprehension. These correlations across comprehension process skills align with the expert comprehension process (Afflerbach et al., 2015), confirming the effectiveness of our framework.

5 Related Work

Comprehension Evaluation Towards the application of LLMs in real-world scenarios, it is necessary to ensure LLMs share a similar comprehension process with experts. However, current comprehension evaluation methods either focus on the stage after the process (matching the answers with human references) (Rajpurkar et al., 2018; Lai et al., 2017; Dua et al., 2019) or the stage before the process (digging the literary skills of LLMs) (Dunietz et al., 2020; Sugawara et al., 2017; Wang et al., 2022; Sugawara et al., 2021). To this end, we propose SCOP to carefully examine the gap between the comprehension processes of LLMs and experts.

Design of Comprehension Tasks Recent years have witnessed plenty of emerging reading comprehension tasks, exploring the comprehension potential of LLMs from different aspects. Some studies focus on challenging LLMs with different document types (Kočiskỳ et al., 2018; Dasigi et al., 2021). They argue that different types of documents need different ways of comprehension, as narrative documents mostly describe events while expository documents mostly explain facts. In line with these studies, we also involve both narrative and expository documents in SCOP. Another part of the literature works on the surface forms of the tasks, including text extraction (Rajpurkar et al., 2018; Dasigi et al., 2019), multi-choice answers (Richardson et al., 2013; Clark et al., 2018), free-from answers (Bajaj et al., 2016), etc. For a full and objective evaluation, we also involve diverse styles of answers but exclude free-form ones, the correctness of whose answers is a matter of preference. There are also studies working on adding "difficulties" to the comprehension burdens. They require LLMs to do causal reasoning (Jin et al., 2024), commonsense reasoning (Zhao et al., 2023), arithmetic calculation (Yuan et al., 2023), etc. We think this is beyond the comprehension process thus we don't involve such studies in SCOP.

6 Conclusion

We propose SCOP to explore the comprehension process of LLMs from a cognitive view. Specifically, it is equipped with a systematical definition of five requisite comprehension skills, a strict data construction framework, and a detailed analysis of various LLMs. Experimental results reveal that all LLMs are still far from achieving the expert-level comprehension process. Further analysis reveals LLMs exhibit inconsistent behavior. They use incorrect comprehension processes despite providing correct answers, highlighting the importance of comprehension processes for reliable models. Additionally, we observe that a better comprehension process can lead to better downstream performance. We hope SCOP can benefit the research on the improvement and deployment of reliable LLMs.

7 Ethic Consideration and Limitations

The testing data are constructed from both existing and newly crawled datasets. For existing datasets, they are all accessed and used in full compliance with their respective licenses and terms of use. Each dataset was reviewed to ensure that the permissible uses under the applicable licenses align with the scope of our research. For newly crawled datasets, we have carefully checked them to make sure they don't contain any personally identifiable information or sensitive personally identifiable information. Therefore, we believe that there is no ethical issue within SCOP.

Part of our testing data is from existing datasets, so some data might already be known to LLMs. In the future, it's important to detect previously seen data and develop better data generation methods. Like other studies on prompting LLMs, our evaluation results may be sensitive to the prompts used. The five skills interact in complex ways, making it difficult to directly confirm the relationships between them. Fortunately, educational research has explored these relationships in depth (Rampey et al., 2009). We believe exploring these in NLP could deepen our understanding of LLMs' comprehension from a cognitive perspective. We focus only on the comprehension process as they are fundamental. The next step is to study how to combine internal knowledge during the thinking process. We believe such thinking is necessary for a better comprehension of LLMs.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (No. 62206191, No. 62272330, No. U24A20328, and No.62306156), in part by the Science Fund for Creative Research Groups of Sichuan Province Natural Science Foundation (No. 2024NSFTD0035), in part by the Natural Science Foundation of Sichuan (No. 2024YFHZ0233), and in part by the Undergraduate Education and Teaching Reform Projects, Nankai University (NKJG2025047).

References

Peter Afflerbach, Byeong-Young Cho, and Jong-Yun Kim. 2015. Conceptualizing and assessing higher-order thinking in reading. *Theory into Practice*, 54(3):203–212.

Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, Vince I Madai, and Precise4Q Consortium. 2020. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, 20:1–9.

Anthropic. 2024. Claude 3.5: A new era of ai.

Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A. Gers, and Alexander Löser. 2019. Sector: A neural model for coherent topic segmentation and classification. *Transactions of the Association for Computational Linguistics*, 7:169–184.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Yevgeni Berzak, Jonathan Malmaud, and Roger Levy. 2020. Starc: Structured annotations for reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5726–5735.

David Caplan and Nancy Hildebrandt. 1988. *Disorders of syntactic comprehension*. MIT Press.

Patricia L Carrell. 1998. Can reading strategies be successfully taught? *Australian review of applied Linguistics*, 21(1):1–20.

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, memory: An archaeology of books known to chatgpt/gpt-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Yiming Cui, Ting Liu, Ziqing Yang, Zhipeng Chen, Wentao Ma, Wanxiang Che, Shijin Wang, and Guoping Hu. 2020. A sentence cloze dataset for chinese machine reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6717–6723.

Pradeep Dasigi, Nelson F Liu, Ana Marasović, Noah A Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5925–5932.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2368–2378.

Jesse Dunietz, Greg Burnham, Akash Bharadwaj, Owen Rambow, Jennifer Chu-Carroll, and Dave Ferrucci. 2020. To test machine comprehension, start by defining comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7839–7859.

Sarah H Eason, Lindsay F Goldberg, Katherine M Young, Megan C Geist, and Laurie E Cutting. 2012. Reader–text interactions: How differential text and question types influence cognitive skills needed for reading comprehension. *Journal of educational psychology*, 104(3):515.

Roger Farr and Robert F Carey. 1986. *Reading: What can be measured?* ERIC.

Norman Frederiksen. 1984. The real test bias: Influences of testing on teaching and learning. *American psychologist*, 39(3):193.

AC Graesser. 2003. What do readers need to learn in order to process coherence relations in narrative and expository text. *Rethinking Reading Comprehension/Guilford Publications*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. 2024. Cladder: A benchmark to assess causal reasoning capabilities of language models.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 252–262.

Walter Kintsch. 1998. *Comprehension: A Paradigm for Cognition*. Cambridge University Press.

Walter Kintsch and Teun A. van Dijk. 1978. Toward a model of text comprehension and production. *Psychological Review*, 85(5):363–394.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Xiang Kong, Varun Gangal, and Eduard Hovy. 2020. Scde: Sentence cloze dataset with high quality distractors from examinations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5668–5683.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 469–473.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*.

DR Krathwohl. 2002. A revision bloom's taxonomy: An overview. *Theory into Practice*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794.

Meta. 2024a. Introducing llama 3.1: Our most capable models to date.

Meta. 2024b. Introducing meta llama 3: The most capable openly available llm to date.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.

OpenAI. 2024a. Gpt-4o mini: Advancing cost-efficient intelligence.

OpenAI. 2024b. Hello, gpt-4o.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. On the risk of misinformation pollution with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1389–1403.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Bobby D Rampey, Gloria S Dion, and Patricia L Donahue. 2009. Naep 2008: Trends in academic progress. nces 2009-479. *National Center for Education Statistics*.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203.

Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241.

Shaurya Rohatgi, Yanxia Qin, Benjamin Aw, Niranjana Unnithan, and Min-Yen Kan. 2023. The acl ocl corpus: Advancing open science in computational linguistics. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10348–10361.

Nancy Nelson Spivey. 1990. Transforming texts: Constructive processes in reading and writing. *Written communication*, 7(2):256–287.

Saku Sugawara, Yusuke Kido, Hikaru Yokono, and Akiko Aizawa. 2017. Evaluation metrics for machine reading comprehension: Prerequisite skills and readability. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 806–817.

Saku Sugawara, Pontus Stenetorp, and Akiko Aizawa. 2021. Benchmarking machine reading comprehension: A psychological perspective. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1592–1612.

Yoshi Suhara and Dimitris Alikaniotis. 2024. Source identification in abstractive summarization. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 212–224.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Paul van den Broek and Christine A Espin. 2012. Connecting cognitive theory and assessment: Measuring individual differences in reading comprehension. *School psychology review*, 41(3):315–325.

Xiaoqiang Wang, Bang Liu, Fangli Xu, Bo Long, Siliang Tang, and Lingfei Wu. 2022. Feeding what you need by understanding what you learned. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 5858–5874.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024b. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.

Ainur Yessenalina, Yisong Yue, and Claire Cardie. 2010. Multi-level structured models for document-level sentiment classification. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1046–1056.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. 2023. How well do large language models perform in arithmetic tasks? *arXiv preprint arXiv:2304.02015*.

Zirui Zhao, Wee Sun Lee, and David Hsu. 2023. Large language models as commonsense knowledge for large-scale task planning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 31967–31987.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 46595–46623.

A Data Collection Details

A.1 Data Statistics

More details about the testing data are shown in Table 5, which includes statistics on document length, the original dataset size, and associated cost.

A.2 Different Document Types

According to Eason et al. (2012), the main types of documents include narrative and expository. Narrative documents are typically structured around events about characters in a temporal sequence. Examples of narrative structures include news articles, stories, and fiction. On the other hand, expository documents focus on presenting facts about a specific topic. Common examples of expository texts are encyclopedias and reports.

A.3 Semantic Similarity Calculating

To select an effective model for calculating semantic similarity, we use the MTEB benchmark (Muennighoff et al., 2023). MTEB tests how well text embeddings perform on different tasks. We focus on four tasks from MTEB that are most relevant for finding semantically similar sentences: Clustering (the task of grouping similar documents together), Pair classification (the task of determining whether two texts are similar), Retrieval (the task of finding relevant documents for a query), and STS (the task of determining how similar two texts are). We add up the models' scores across these tasks and pick the one with the highest overall score. We prefer a smaller-scale model(less than 500M parameters) because it often performs better in sentence-pair similarity tasks and reduces computational costs. Finally, we choose stella-en-400M-v5[10] for calculating similarity using cosine similarity scores.

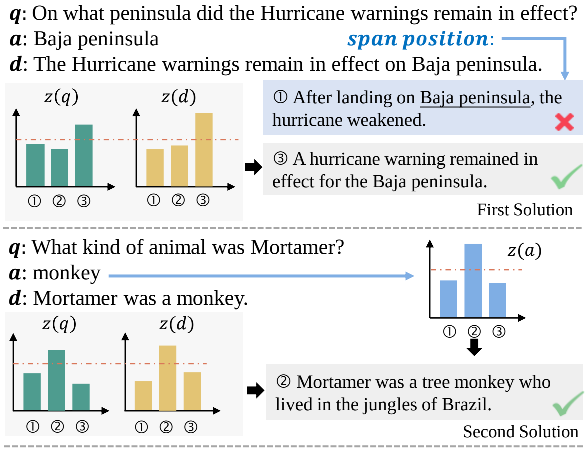
q: On what peninsula did the Hurricane warnings remain in effect?
a: Baja peninsula
span position:
d: The Hurricane warnings remain in effect on Baja peninsula.
z(q)
z(d)
① ② ③
① ② ③
① After landing on Baja peninsula, the hurricane weakened.
③ A hurricane warning remained in effect for the Baja peninsula.
First Solution
q: What kind of animal was Mortamer?
a: monkey
d: Mortamer was a monkey.
z(a)
① ② ③
z(q)
z(d)
① ② ③
① ② ③
② Mortamer was a tree monkey who lived in the jungles of Brazil.
Second Solution

Figure 5: The two solutions for identifying locating-type questions. The first solution is used to handle datasets with answer annotations, while the second handles datasets without them.

A.4 Details of Locating Questions Collection

When identifying locating questions, we retain data with the document including more than 4 sentences across the three datasets. As LLMs are required to identify supporting sentence, fewer sentences would impact evaluation results. For MCTest, we also exclude data labeled "multiple", meaning a question can be supported by multiple sentences. After preprocessing, we use two solutions described in Section 3.1 (illustrated in Figure 5) to filter locating questions. For the second solution, We observe that $S(*, \star)$ doesn't work well when the answer is too short to be well-encoded in embedding spaces. As such, the pseudo-supporting sentence is retrieved by BM25 (Robertson and Walker, 1994). Finally, the first solution enables us to filter about 40% of questions with one supporting sentence in SQuAD 2.0 and over 50% in NewsQA. The second solution filters about 40% locating questions in MCTest.

Impact of the Answer on Retrieval When the answer to a question appears only once in the context, using only the declarative sentence for re-

[10] https://huggingface.co/dunzhang/stella_en_400M_v5

Skill	Source	Mean Tokens	Max Tokens	Original Size	Size	Cost (dollars)
Locating	SQuAD v2.0	187	714	5928	479 (8.08%)	0.72
	NewsQA	735	2463	10292	984 (9.56%)	0.89
	MCTest	253	514	1160	72 (6.21%)	0.11
Inferring	HotpotQA	1329	3035	7405	604 (8.16%)	0.69
	MusiQue	2413	4449	2417	510 (21.10%)	0.40
	RACE	338	974	9821	547 (5.57%)	1.25
Connecting	SCDE	276	1010	625	625 (100.00%)	-
	ACL OCL	514	1104	-	169 (-)	-
Organizing	ClimateCentral	1046	3596	-	108 (-)	-
	WikiHow	1459	4583	-	146 (-)	-
Selecting	SourceSum	294	542	143	143 (100.00%)	-
	ACL OCL	535	1490	-	295 (-)	-

Table 5: Statistics on document length, size, and associated cost of the testing data.

Model	Locating			Inferring			Connecting		Orgnazing		Selecting	
	MCTest	SQuAD	NewsQA	HotpotQA	MusiQue	RACE	SCDE	ACL	Climate	WikiHow	SourceSum	ACL
Deepseek-R1	97.22	99.16	86.89	51.32	36.67	44.97	72.16	40.83	22.22	34.25	13.99	10.17
GPT-4o-mini	91.67	97.49	85.87	34.44	11.37	37.84	68.25	10.06	12.04	30.82	10.49	12.54

Table 6: Performance of additional LLMs' comprehension process across five skills.

Document: In April 1191 Richard the Lion-hearted left Messina with a large fleet in order to reach Acre. But a storm dispersed the fleet. After some searching, it was discovered that the boat carrying his sister and his fiancée Berengaria was anchored on the south coast of Cyprus.

Question: What ruined Richard's plans to reach Acre?

Answer: a storm

Table 7: An example where the answer appears only once in the context. The annotated span is in bold, and the annotated supporting sentence is underlined.

trieval would be impacted. Taking Table 7 as an example, the annotated supporting sentence would be retrieved using the declarative sentence as the query. This question would then be classified as a locating question. However, this annotated supporting sentence does not provide sufficient information to answer the question. To address this issue, we use both the declarative sentence and the question to retrieve. The final set of supporting sentence candidates is the intersection of these retrieval results. This method ensures that the retrieved sentence contain information relevant to both the question and the answer.

A.5 Details of Inferring Questions Collection

We use the pretrained model en_core_web_trf[11] in SpaCy to perform constituency parsing.

Some subquestion candidates cannot be subquestions even if they are noun phrases with named entities. For example, in the question "Who founded the company that distributed the film UHF?", the candidate "the film UHF" is not a subquestion because it does not target anything. In contrast, in the question "Who is the partner of Green performer?", the candidate "the Green performer" is a valid subquestion becuase of targeting "Who is the performer of the Green". These candidates are always found at the deepest edges of the syntactic trees. We check their validity by turning them into yes-no questions. For example, "the film UHF" is transformed into "Is UHF the film" and "the Green performer" is transformed into "Is Green performer?" through template filling. We then ask Llama-3-8B-Instruct to answer these questions with the corresponding documents. If the answer is "No", this candidate is confirmed as a subquestion.

After filtering out invalid candidates, we remove those candidates targeting the same fact. We use kendall's tau coefficient to measure the correlation between two similarity distribution. This coefficient is useful for data considering rankings, which helps us focus on the similarity rankings between candidates and document sentences. The part-of-speech used to identify inferring questions from comparison questions is performed by SpaCy.

A.6 Details of Connecting Data Collection

Noisy papers with messy code from ACL OCL corpus are filtered using regular expression. Besides, we only retain papers published at EMNLP and ACL conferences because these papers have a unified structure. This ensures data consistency.

[11] https://spacy.io/models/en

A.7 Details of Organizing Data Collection

We use the Python library Scrapy[12] to develop a crawler for extracting documents from the Climate-Central website. We retain documents with more than two subheadings, ensuring that LLMs would need to segment the documents into at least three subsections. For the Wikihow dataset, we filter out documents that contain too many short sentences (fewer than 5 words), as such documents are more like procedural text rather than expository text.

A.8 Details of Human Evaluation

To assess the validity of our testing data, we randomly pick 50 samples from each dataset. We instruct three workers with English level certificates to label whether the given sample meets the definition of the specific tasks. If all workers agree that the sample is valid, it is scored 1, otherwise 0. The average validity score is calculated by summing these scores and dividing by the total number of samples. As a result, the validity score is 0.81 with an inter-annotator agreement percentage of 73%, indicating a high reliability of our testing data.

B Additional Experiments

B.1 The Influence of Memorized Data on Performance

In this pilot experiment, we explore how memorized data affects the evaluation performance of LLM. We classify questions that can be directly answered by an LLM without any context as "memorized data", while questions requiring reasoning or additional context are categorized as "non-memorized data". Using this distinction, we evaluate the performance of Llama 3.1-70B on both types of data. The results, shown in Table 8, indicate that the model performs better on memorized data compared to non-memorized data. To reduce the influence of data contamination on our evaluation, we filter out memorized data in Section 3.6.

	HotpotQA	MusiQue	RACE
non-memorized	62.58	30.98	46.07
memorized	62.75	32.93	49.17

Table 8: Performance of Llama3.1-70B on memorized vs. non-memorized data.

Document:
...(24) There are efforts underway to similarly capture and use methane from agriculture and waste facilities known as biogas. (26) There is a growing range of strategies to otherwise reduce methane production from farming and landfills, including alternative feed for cows and composting to reduce waste. (27) Climate Central's full report on methane elaborates on sector-specific strategies and key initiatives to reduce methane from these main sources.
(28) Accurate emissions information is key for setting priorities and tracking progress–or lack thereof–toward reduction goals. (29) But collecting data on an odorless, invisible gas requires specialized technology, so direct measurements of methane emissions are thin. ...

Subheading: How methane is measured

Table 9: A failure case where GPT-4o fails to organize. We add a line break at sentence (28) for clarity. The original document does not have this break.

B.2 Exploring Additional LLMs with SCOP

In Table 6, we observe Deepseek-R1 and GPT-4o-mini perform consistently with our findings obtained from SCOP. Deepseek-R1, with its strong reasoning capability, outperforms most LLMs in the inferring skill. In a pilot study, we also test Llama-3.1-7B but exclude it from the main evaluation due to its extremely poor performance (e.g., 0.59% in connecting, 7.4% in organizing, and 3.4% in selecting).

B.3 Case Study

We present a failure example from the climate dataset for the organizing skill using GPT-4o, shown in Table 9. This task requires LLMs to determine where subheadings should be placed within a document. In this case, the correct position for the subheading is clearly (28). However, GPT-4o incorrectly place it at (29). This mistake may be due to the overlapping token "measure" in both the subheading and sentence (29). This indicates LLMs may not follow a correct comprehension process but rather rely on shortcuts.

We use an example from the RACE dataset to explain why we do not directly treat the sentences retrieved by LLM as the golden supporting sentences. Using the sample Table 10 as an example, the LLM selects sentences 8, 11, 12, and 13 as the

[12] https://scrapy.org/

Document sentences:

(1) Long, long ago there was an old man.
(2) He had a very big orange tree in his garden.
(3) On the tree there were many fine oranges.
(4) One day the old man found one of the oranges was bigger than the others.
(5) It was as big as a watermelon.
(6) So he took the big orange to the king.
(7) The king was very happy and gave the old man a lot of money for it.
(8) When a rich man heard of this, he said to himself, "It is only an orange."
(9) "Why did the king give him so much money?",
(10) "If I take my gold cup to the king, he will give me much more money for it.",
(11) The next day when the king got the gold cup, he said to the rich man, "What a beautiful cup!"
(12) "I'll give you something for it."
(13) "Please take the great orange."

Questions: Was the rich man very happy at last?

Table 10: A sample from RACE.

supporting sentences. However, treating these sentences as the golden results and using this sample as an inference question conflicts with our definition of the comprehension process. This is because answering the question requires knowledge about human emotions: a person feels happy when they achieve their goal. This kind of knowledge is not explicitly stated in the document. To address this, we also extract supporting sentences based on semantic similarity, resulting in sentences 7, 8, and 9. Since fewer than two of these sentences overlap with the ones chosen by the LLM, we exclude this sample from the inference questions. This case shows that the supporting sentences extracted by LLMs are influenced by their background knowledge. Therefore, we incorporate semantic similarity to ensure that the test data more accurately matches the required definition.

C Data Examples

C.1 Locating

Document: One of the first Norman mercenaries to serve as a Byzantine general was Hervé in the 1050s. By then however, there were already Norman mercenaries serving as far away as Trebizond and Georgia. They were based at Malatya and Edessa, under the Byzantine duke of Antioch, Isaac Komnenos. In the 1060s, Robert Crispin led the Normans of Edessa against the Turks. Roussel de Bailleul even tried to carve out an independent state in Asia Minor with support from the local population, but he was stopped by the Byzantine general Alexius Komnenos.

Question: When did Hervé serve as a Byzantine general?
Answer: in the 1050s
Supporting sentence: One of the first Norman mercenaries to serve as a Byzantine general was Hervé in the 1050s.

Table 11: Example of a locating question in SQuAD 2.0

Document: Authorities have recovered 54 bodies after a ferry crammed with people capsized in southern Bangladesh , police said Sunday . Among the victims were 22 children and 15 women , said Nazrul Islam , the police chief of Bhola district where the accident occurred Friday . Thirty more passengers are believed missing and presumed dead , he said . “ Hopefully , in few hours , we should be able to confirm the exact number of missing -LRB- people -RRB- , ” Islam said . The boat had a capacity of 1,500 but was overcrowded with about 2,000 people who were traveling from the capital , Dhaka , to their homes in Bhola for the Muslim festival of Eid al-Adha . The boat toppled as passengers weighted down one side to disembark , Islam said . Police and firefighters rushed to aid passengers , many of whom were trapped in the lower deck . CNN ’s Harmeet Shah Singh contributed to this report.

Question: what was traveling ?

Answer: 2,000 people

Supporting sentence: The boat had a capacity of 1,500 but was overcrowded with about 2,000 people who were traveling from the capital , Dhaka , to their homes in Bhola for the Muslim festival of Eid al-Adha .

Table 12: Example of a locating question in NewsQA

Document: Jenny was standing on a rock. Suddenly, she had to sneeze. After she sneezed, she walked away. She finally got to the park and saw her daddy. Her daddy gave her some milk. Jenny drank the milk in a big hurry. She loved milk. She walked over and turned a switch. She walked to the lake. Jenny was in a big hurry and went really fast. She got to the lake and sat down. Jenny began thinking. Jenny wanted to go on a trip to Florida. Jenny did not want to go someplace cold. Jenny did not want to go to the moon. Jenny did not want to go to France. Jenny stood up to fold her towel. She never folded her shirts or pants. Jenny would start her art for her aunt in a few hours. She knew she would use a lot of time making that art. Her aunt would love the art.

Question: Where did Jenny want to go on a trip to?

Options: A. Florida B. someplace cold C. France D. the moon

Answer: A

Supporting sentence: Jenny wanted to go on a trip to Florida.

Table 13: Example of a locating question in MCTest

C.2 Inferring

Document:
I work at a university in the USA. There, my team and I are trying to learn more about the American black duck, a kind of water bird. And now we are using an exciting piece of equipment called a “night vision scope”. By using it, we can see the ducks in the dark. We’re worried about black ducks mainly because their numbers are falling _____________. And we don’t know whether there’s enough food on the east coast for these birds. There’s lots of information about their daytime activities, but nothing about what they do at night, because we don’t have the equipment. But this new “scope” will make really clear pictures, even on moonless nights, so we will be able to find out more about the ducks.It is very hard work. There are four of us, and we each work six hours every day. We study ducks in different places, and I sometimes have to take a boat to where I need to work. The weather is not helpful because most of the time it’s wet...

Question: What does the writer’ team hope to find out about American black ducks?
Options:
A. What food they feed on.   B. What makes the east coast a good place for them.
C. What they do at night.   D. What animals like to stay with them.
Answer: C
Supporting sentence:
1. There’s lots of information about their daytime activities, but nothing about what they do at night,because we don’t have the equipment.
2. But this new “scope” will make really clear pictures, even on moonless nights, so we will be able to find out more about the ducks.

Table 14: Example of an inferring question in RACE

Document:
Klaus Meine || Klaus Meine (born 25 May 1948) is a German vocalist, best known as the lead singer of the hard rock band Scorpions. He and guitarist Rudolf Schenker are the only two members of the group to appear on every Scorpions album. Meine was placed at #22 on Hit Parader’s Top Heavy Metal Vocalists of All Time list in 2006.
A Moment in Chiros || A Moment in Chiros is American heavy metal vocalist Lance King’s studio debut album as a solo artist, featuring the musical contributions of many of his friends, contemporaries, and business associates.
Geoff Tate || Geoff Tate (born Jeffrey Wayne Tate, January 14, 1959; he later changed his first name to Geoffrey) is a German-born American singer and musician. He rose to fame with the progressive metal band Queensrÿche, who had commercial success with their 1988 album “Operation: Mindcrime” and 1990 album “Empire”. Tate is ranked fourteenth on Hit Parader’s list of the 100 Greatest Metal Vocalists of All Time. He was voted No. 2 on “That Metal Show”’s top 5 hard rock vocalists of the 1980s. In 2012, he won the Vegas Rocks! Magazine Music Award for “Voice in Progressive Heavy Metal”. In 2015, he placed ninth on OC Weekly’s list of the 10 Best High-Pitched Metal Singers. After his farewell tour as Queensrÿche, he renamed his band Operation: Mindcrime, after the Queensrÿche album.
Dee Snider || Daniel “Dee” Snider (born March 15, 1955) is an American singer-songwriter, screenwriter, radio personality, and actor. Snider came to prominence in the early 1980s as lead singer of the heavy metal band Twisted Sister. He was ranked 83 in Hit Parader's Top 100 Metal Vocalists of All Time.
Lance King || Lance King (born November 23, 1962) is an American heavy metal vocalist specializing in melodic rock, progressive, and power metal. Lance has sung with many groups over the last 35 years and started the record label Nightmare in 1990 to release his own music. He is presently still at the helm of the label.
Avian (band) || Avian is a melodic power metal band founded in 2002 by guitarist Yan Leviathan. The band features singer Lance King. In 2005, they released their debut album “From the Depths of Time”, a concept album dealing with the end of days and a warning to mankind. Musically, Avian is influenced by bands such as Iron Maiden, HammerFall, Savatage, and Megadeth. In December 2006, Avian was an opening act for Twisted Sister. Their second album, titled “Ashes and Madness”, was released in September 2008. In early 2010, Lance decided to leave the band to focus on family and professional obligations and was replaced with Brian Hollenbeck, who appeared on their first EP, entitled “The Path”, which was released in September 2010.
Sully Erna || Salvatore Paul “Sully” Erna Jr. (born February 7, 1968) is the American vocalist and guitarist for the American hard rock band Godsmack. He is also a harmonica player, percussionist, and pianist, performing these on albums and at live shows. He was ranked 47th in the Top 100 Heavy Metal Vocalists by Hit Parader.
Han Seung-yeon || Han Seung-yeon (born July 24, 1988), better known mononymously as Seungyeon, is a South Korean singer and actress. She is best known as the former main vocalist of the South Korean girl group Kara.
...

Question: Are Lance King and Han Seung-yeon both heavy metal vocalists?
Answer: No
Supporting sentence:
1. Lance King (born November 23, 1962) is an American heavy metal vocalist specializing in melodic rock progressive and power metal.
2. She is best known as former main vocalist of the South Korean girl group Kara.

Table 15: Example of an inferring question in HotpotQA

Document:
West DeLand, Florida || West DeLand is a census-designated place (CDP) in Volusia County, Florida, United States. The population was 3,535 at the 2010 census.
Kendall Green, Pompano Beach, Florida || Kendall Green was a census-designated place (CDP) in Broward County, Florida, United States, and is now a neighborhood of Pompano Beach, Florida. The population was 3,084 at the 2000 census.
Ridgecrest, Florida || Ridgecrest is a census-designated place (CDP) in Pinellas County, Florida, United States. The population was 2,558 at the 2010 census.
Wade Hampton, South Carolina || Wade Hampton is a census-designated place (CDP) in Greenville County, South Carolina, United States. The population was 20,622 at the 2010 census. It is named for American Civil War general and South Carolina governor Wade Hampton.
Zephyrhills North, Florida || Zephyrhills North is a census-designated place (CDP) in Pasco County, Florida, United States. The population was 2,544 at the 2000 census.
Tamiami, Florida || Tamiami is a census-designated place (CDP) in Miami-Dade County, Florida, United States. The population was 55,271 at the 2010 census.
Hampton Double Square Historic District || The Hampton Double Square Historic District is a historic district located in Hampton, Iowa, United States. It has been listed on the National Register of Historic Places since 2003. At the time of its nomination it contained 43 resources, which included 28 contributing buildings, two contributing sites, 10 non-contributing buildings, one non-contributing site, one non-contributing structures, and one non-contributing object. The town of Hampton was laid out by H.P. Allen, who was the county surveyor, in June 1856. The original plat was eight blocks by eight blocks in the shape of an "L". Near the center of the "L" was the two-block, or double, square. While many county seats in Iowa have a courthouse square, the double square is a rarity. Four double squares were platted in Iowa, but only those in Hampton and Sidney survived their early period of development. Estherville's square was platted as a four-block square, but its development created a double square instead. Hampton has the only symmetrical double square plan in the state. The double square exemplifies the two primary functions of a public square, both commercial and public development.
Sean Hampton || Born in Ocala, Florida, Hampton is the youngest of five children of a dentist father and a professional model mother. After graduating high school, Hampton enrolled at Stetson University to pursue a career in law. While in school he not only joined Sigma Nu fraternity (Delta Mu chapter), but caught onto acting. After college he married his current wife Jennifer and the two moved to Los Angeles where they currently reside.
Gladeview, Florida || Gladeview is a census-designated place (CDP) in Miami-Dade County, Florida, United States. The population was 11,535 at the 2010 census.
Solana, Florida || Solana is an unincorporated community and census-designated place (CDP) in Charlotte County, Florida, United States. The population was 742 at the 2010 census. It is part of the Punta Gorda, Florida Metropolitan Statistical Area.
Ocala, Florida || Ocala is a city located in Northern Florida. As of the 2013 census, its population, estimated by the United States Census Bureau, was 57,468, making it the 45th most populated city in Florida.
...

Question: Where is Sean Hampton's birth place in the state of Florida?
Answer: in Northern Florida
Supporting sentence:
1. Born in Ocala, Florida, Hampton is the youngest of five children of a dentist father and a professional model mother.
2. Ocala is a city located in Northern Florida. As of the 2013 census, its population, estimated by the United States Census Bureau, was 57,468, making it the 45th most populated city in Florida.

Table 16: Example of an inferring question in MusiQue

C.3 Connecting

Document: Confidence is very important in daily life . It can help you to develop a healthy attitude . But how to be more confident ? Here are some suggestions : <1> If you like singing , sing as much as you can . In some ways , a hobby can make you outstanding . And it will make you happy and confident . <2> Exercise makes you tired but relaxed . A strong body helps you be full of confidence . <3> Fear comes along with failure . But it 's easy to overcome if you know that failure is part of your life . Try to start again and believe you can do better . <4> When you are not confident , you will speak in a low voice . Try to speak loudly enough so that people can hear you clearly . The high voice can help you become more confident . <5> Write down a list of things you did during the day to see how many things you have done well .
Candidates:
1. Play sports .
2. Pick up a hobby .
3. Speak loudly .
4. Get rid of fear .
5. Ask for help .
6. Find your advantages
Answer: 2, 1, 4, 3, 6

Table 17: Example of a connecting data in SCDE

Document: Text segmentation is a traditional NLP task that breaks up text into constituents, according to predefined requirements. It can be applied to documents, in which case the objective is to create logically coherent sub-document units. <1> This task is often referred to as document segmentation or sometimes simply text segmentation . <2> Documents are often multi-modal, in that they cover multiple aspects and topics; breaking a document into uni-modal segments can help improve and/or speed up down stream applications. For example, document segmentation has been shown to improve information retrieval by indexing sub-document units instead of full documents. Other applications such as summarization and information extraction can also benefit from text segmentation <3> breaks up pieces of text into sub-sentence elements called Elementary Discourse Units (EDUs). EDUs are the minimal units in discourse analysis according to the Rhetorical Structure Theory. In Figure 2 we show examples of EDU segmentations of sentences. For example, the sentence "Annuities are rarely a good idea at the age 35 because of withdrawal restrictions" decom-poses into the following two EDUs: "Annuities are rarely a good idea at the age 35" and "because of withdrawal restrictions", the first one being a state-ment and the second one being a justification in the discourse analysis. In addition to being a key step in discourse analysis, discourse segmentation has been shown to improve a number of downstream tasks, such as text summarization, by helping to identify fine-grained sub-sentence units that may have different levels of importance when creating a summary. <4> Koshorek et al. proposed the use of 4708 hierarchical Bi-LSTMs for document segmentation. Simultaneously, Li et al. introduced an attention-based model for both document seg-mentation and discourse segmentation, and Wang et al. obtained state of the art results on dis-course segmentation using pretrained contextual embeddings. Also, a new large-scale dataset for document segmentation based on Wikipedia was introduced by Koshorek et al., providing a much more realistic setup for evaluation than the previously used small scale and often synthetic datasets such as the Choi dataset. However, these approaches are evaluated on different datasets and as such have not been compared against one another. Furthermore they mostly rely on RNNs instead of the more recent transformers and in most cases do not make use of contextual embeddings which have been shown to help in many classical NLP tasks. <5> 1. We compare recent approaches that were pro-posed independently for text and/or discourse segmentation on three public datasets. 2. We introduce three new model architectures based on transformers and BERT-style con-textual embeddings to the document and dis-course segmentation tasks. We analyze the strengths and weaknesses of each architecture and establish a new state-of-the-art. 3. We show that a simple paradigm argued for by some of the earliest text segmentation algorithms can achieve competitive performance in the current neural era. 4. We conduct ablation studies analyzing the im-portance of context size and model size.

Candidates:

1. Our experiments showed that all of our models improve the current state-of-the-art.
2. These units, or segments, can be any structure of interest, such as paragraphs or sections.
3. In this work we aim at addressing these limitations and make the following contributions:
4. A related task called discourse segmentation
5. Naturally these results do not imply that hierarchical models should be disregarded.
6. In Figure 1 we show one ex-ample of document segmentation from Wikipedia, on which the task is typically evaluated
7. Multiple neural approaches have been recently proposed for document and discourse segmentation.

Answer: 2, 6, 4, 7, 3

Table 18: Example of a locating question in ACL OCL

C.4 Organizing

Document: (0) This summer's relentless record heat has stuck around into fall. (1) The planet just had a record-shattering September—the seventh-warmest for the U.S. (2) This is bad news for the 50 million people in the U.S. with allergies to ragweed pollen in the late summer and early fall. (3) In most U.S. areas, ragweed pollen typically peaks in September and lasts through October. (4) But warmer fall temperatures extend the ragweed growing season. (5) Ragweed, which is found in most U.S. states, is the main cause of fall allergies. (6) A single ragweed plant can produce up to 1 billion pollen grains that are carried by wind and cause a range of symptoms. (7) Ragweed can also thrive in both rural and urban areas. (8) A 2003 study suggests that the urban heat island effect can even help ragweed grow faster and produce more pollen in cities. (9) See Urban Heat Hot Spots to find the strongest urban heat islands within your city. (10) How is fall warming affecting local fall allergy seasons? (11) Climate Central analysis explores this question. (12) Studies have found that the length of ragweed pollen season across the U.S.—from Texas to North Dakota—is strongly linked with the number of fall days until the first frost. (13) Climate Central therefore assessed how the number of consecutive freeze-free days (with minimum temperatures above 32°F) during the fall season (Sept-Nov) has changed since 1970 in 201 U.S. cities. (14) The freeze-free fall season lengthened in 164 cities, or 82% of the 201 analyzed. (15) Across these 164 cities, the freeze-free fall season lengthened by 11 days on average. (16) The freeze-free fall season is now at least two weeks longer in 53 cities. (17) Over half (58%) of these 53 cities were in the Northeast, Upper Midwest, and Northwest—consistent with research finding that ragweed pollen season has grown fastest at higher latitudes. (18) The four cities where the fall freeze-free season has grown the most since 1970 are: Reno, Nev. (+39 days); Bend, Ore. (+33 days); Toledo, Ohio (+28 days); Boise, Idaho (+27 days). (19) The widespread increase in freeze-free fall days can prolong allergy-inducing pollen production by the 17 types of ragweed that grow across the U.S. during the late summer and fall. (20) Seasonal allergies can already last from early spring through late fall. (21) But warming from carbon pollution results in more freeze-free days each year, giving plants more time to grow and release allergy-inducing pollen...

Subheadings:	**Answer:**
Summer heat lingering into fall	Summer heat lingering into fall, 0
Growing season lasting later into fall	Growing season lasting later into fall, 7
Warming climate, longer pollen season, worse allergies	Warming climate, longer pollen season, worse allergies, 15
Mold can cause fall allergies, too	Mold can cause fall allergies, too, 21

Table 19: Example of an organizing data in ClimateCentral

Document: (0) Look for auctions held in less popular or crowded areas. (1) Like any auction, the more crowded it is, the more competition you may have. (2) A big crowd could drive the bidding prices up or cause you to lose out on a bid for a vehicle. (3) Look for auctions that are situated in less populated areas or tend to fly under the radar. (4) You can search for police auctions in certain areas online. (5) Focus on auctions outside of a major city, if possible, or in a smaller town or city, as these may be less crowded than auctions held in larger cities or known areas. (6) Research the vehicles listed online a few days before the auction date. (7) Most auctions will list the vehicles that will be available at the auction a few days before the auction date. (8) Look over each listing and identify which vehicles you are interested in bidding on. (9) You should try to choose at least one to two vehicles in the event you lose out on a bid so you have a backup vehicle you can still bid on. (10) If you have your eye on a Mercedes-Benz CLK listed online, for example, you should note the details listed for the car. (11) Then, you should research the market value of a used Mercedes-Benz CLK and determine how much you would be willing to bid for the car. (12) Make sure you are clear on the maximum you would be willing to spend on the car as this can prevent you from overbidding in the chaos of the auction. (13) Bring cash or proof of an approved loan to the auction. (14) Police auctions will only take payment in cash or proof of an approved loan for the winning bid. (15) If you are planning to pay with an approved loan from your bank, you will need to be able to cover a minimum deposit for the full cost of the vehicle. (16) You will also need to cover the cost of taxes, title, and registration fees. (17) Cars sold at auction do not come with a warranty and are considered “as is” so you will likely need to purchase insurance and a warranty for the car once you buy it. (18) You will also need enough money to cover the cost of towing the car from the auction and the cost of cutting new keys for the vehicle if it is sold without keys. (19) Take a set of tools, car oil, and an air pressure gauge. (20) You will not be able to drive the vehicles before you bid on them so inspecting the car beforehand with tools, car oil, and an air pressure gauge can help to ensure the car is in working order. (21) Show up early and check in. (22) The vehicles at a police auction are often shown in a set order so get to the auction early and check in with the auction. (23) You can get a copy of the showing list at check in and have a chance to inspect the vehicles you are interested in before the auction starts. (24) Inspect the vehicles you are interested in bidding on. (25) Use your set of tools to do a quick inspection of the vehicles you plan to bid on. (26) The vehicles appear at the auction untouched, which means they are in the exact state they were in when they were confiscated by the police. (27) Be prepared for the vehicles to be filthy, damaged, or full of someone else’s stuff. (28) Do not be put off by surface level dirt or strong smells, as these can be cleaned out as long the vehicle’s parts are in good shape. (29) Lift the hood of the vehicle and give it a good inspection. (30) Look at the brakes, the shocks, and the quality of the tires on the vehicle. (31) This will help you determine if the vehicle is worth bidding on and how much you should bid for the vehicle. (32) Do not bid more than you can afford. (33) It can be easy to get caught up in the chaos of bidding wars and quick sales at the auction, so focus on staying calm and not bidding more than you can afford. (34) Remember your predetermined limit you set for yourself as you bid on the vehicles you are interested in and try not to overbid in an attempt to outbid someone else. (35) Avoid making quick, in the heat of the moment decisions and really be certain you want a vehicle before you start bidding on it. (36) You don’t want to end up having to pay more for a vehicle than you can afford or than it’s worth because you got caught up in a bidding war. (37) Check if there is a towing company on site. (38)...

Subheadings:	**Answer:**
Summer heat lingering into fall	Summer heat lingering into fall, 0
Attending the Police Impound Auction	Attending the Police Impound Auction, 21
Taking Your Car Home	Taking Your Car Home, 38

Table 20: Example of an organizing data in Wikihow

C.5 Selecting

Document: Not everyone should be behind the wheel of a $50,000 car. That’s one lesson to take away from a video posted by YouTube user Richard Stewart showing a Porsche Cayman flying out of control as it speeds from a green light on Prince Edward Island in Canada. The sports car swerves wildly before smashing into the concrete median. A wheel even comes off before the car finally comes to a halt. “Just cause you have a nice car doesn’t make you a good driver. Don’t let your son drive your Porsche!” wrote Stewart on YouTube about the crash. KHOU reports that police have not made the identity of the driver public but have said that a 31-year-old driver was cited for the crash, leaving the car totaled as it was towed away. The footage begins with the Porsche idling at a green light. The car booms ahead at a dangerous speed. Almost immediately the driver begins to lose control. The unidentified man veers wildly across the dividing line. The car is twisting at such dangerous speeds a wheel comes loose. Finally, the car comes to a halt, a total wreck waiting for the tow truck.

Key sentences:

1. That’s one lesson to take away from a video posted by YouTube user Richard Stewart showing a Porsche Cayman flying out of control as it speeds from a green light on Prince Edward Island in Canada.
2. KHOU reports that police have not made the identity of the driver public but have said that a 31-year-old driver was cited for the crash, leaving the car a totaled as it was towed away.
3. Finally the car comes to a halt, a total wreck waiting for the tow truck .

Table 21: Example of a selecting data in sourcesum

Document: Sentiment is personal; the same sentiment can be expressed in various ways and the same expression might carry distinct polarities across different individuals. Current mainstream solutions of sentiment analysis overlook this fact by focusing on population-level models. However, only one global model is estimated there, and the details of how individual users express diverse opinions cannot be captured. More importantly, existing solutions build static sentiment models on historic data; but the means in which a user expresses his/her opinion is changing over time. To capture temporal dynamics in a user's opinions with existing solutions, repeated model reconstruction is unavoidable, albeit it is prohibitively expensive. As a result, personalized sentiment analysis requires effective exploitation of users' own opinionated data and efficient execution of model updates across all users. To address these challenges, we propose to build personalized sentiment classification models via shared model adaptation. Our solution roots in the social psychology theories about humans' dispositional tendencies. Humans' behaviors are shaped by social norms, a set of socially shared "feelings" and "display rules" about how one should feel and express opinions. Intuitively, personalized model adaptations can be considered as a set of related tasks in individual users, which contribute to a shared global model adaptation. In particular, we assume the distinct ways in which users express their opinions can be characterized by a linear classifier's parameters, i.e., the weights of textual features. Personalized models are thus achieved via a series of linear transformations over a globally shared classifier's parameters, e.g., shifting and scaling the weight vector. This globally shared classifier itself is obtained via another set of linear transformations over a given base classifier, which can be estimated from an isolated collection beforehand and serves as a prior for shared sentiment classification. The shared global model adaptation makes personalized model estimation no longer independent, such that regularity is formed across individualized learning tasks. We empirically evaluated the proposed solution on two large collections of reviews, i.e., Amazon and Yelp reviews. Extensive experiment results confirm its effectiveness: the proposed method outperformed user-independent classification methods, several state-of-the-art model adaptation methods, and multi-task learning algorithms.

Key sentences:

1. As a result, personalized sentiment analysis requires effective exploitation of users' own opinionated data and efficient execution of model updates across all users.
2. To address these challenges, we propose to build personalized sentiment classification models via shared model adaptation.
3. The shared global model adaptation makes personalized model estimation no longer independent, such that regularity is formed across individualized learning tasks.
4. We empirically evaluated the proposed solution on two large collections of reviews, i.e., Amazon and Yelp reviews.

Table 22: Example of a selecting data in ACL OCL

D Prompts for five skill evaluation

D.1 Locating

Type	Prompt
EQA	Answer the provided question based on the given context. First, identify the sentence that supports the answer to the question, then output both the supporting sentence and the answer in JSON format, ensure the answer is directly extracted from the original text. Response with the JSON only! Here is an example: #Context ##context## # Question When was the Duchy of Normandy founded? # Output { "supporting_sentence": "The Duchy of Normandy, which began in 911 as a fiefdom, was established by the treaty of Saint-Clair-sur-Epte between King Charles III of West Francia and the famed Viking ruler Rollo, and was situated in the former Frankish kingdom of Neustria.", "answer": "911" } # Context ##context## # Question ##question## # Output
MCQA	Given a multiple-choice question, select the correct option based on the provided context. To complete the task, first identify the sentence that supports the answer, and then output both the supporting sentence and the chosen option in JSON format. Response with the JSON only! Here is an example: # Context ##context## # Question What broke the fence? # Options ["A tree.", "A raccoon.", "John.", "The things that were missing from the back yard."] # Output { "supporting_sentence": "A tree, weighted down by the snow, had fallen on the fence on a windy day and broken a section." , "answer": "A tree." } # Context ##context## # Question ##question## # Options ##options## # Output

Table 23: Prompts for locating skill evaluation.

D.2 Inferring

Type	Prompt
EQA	Answer the provided question based on the given context. First, identify relevant sentences from the context that support to answer the question. Then, integrate these sentences to form an answer, ensure the answer is directly extracted from the original text. Output both the identified sentences and the final answer in JSON format. Use <S> to separate different supporting sentences. Response with the JSON only! Here is an example: # Context ##context## # Question Were Scott Derrickson and Ed Wood of the same nationality? # Output { "supporting_sentence": "Ed Wood is a 1994 American biographical period comedy-drama film directed and produced by Tim Burton, and starring Johnny Depp as cult filmmaker Ed Wood. <S> Scott Derrickson (born July 16, 1966) is an American director, screenwriter and producer.", "answer": "yes"} # Context ##context## # Question ##question## # Output
MCQA	Given a multiple-choice question, select the correct option based on the provided context. First, identify relevant sentences from the context that support to answering the question. Then, integrate these sentences to determine the correct option. Output both the identified supporting sentences and the final selected option in JSON format. Use <S> to separate different supporting sentences. Response with the JSON only! Here is an example: # Context ##context## # Question Some people in the Australian outback can't get to a doctor quickly. Because _ # Options ["there are few doctors there", "the nearest doctor is sometimes very far away from them", "there is always heavy traffic on the road", "they don't want to see a doctor"] # Output { "supporting_sentence": "But people in the Australian outback can't get to a doctor quickly.<S> The nearest doctor is sometimes hundreds of kilometers away so they have to call him on a two-way radio.", "answer": "the nearest doctor is sometimes very far away from them" } # Context ##context## # Question ##question## # Options ##options## # Output

Table 24: Prompts for inferring skill evaluation.

D.3 Connecting

Prompt

Complete the following passage by selecting and inserting sentences from the provided candidates into the designated blanks, ensuring the passage is coherent and logically structured. The blanks within the passage are denoted by <NUM>. Output the chosen sentences in the sequence corresponding to the blanks in the passage in JSON format. Response with the JSON only!
Here is an example:
#passage:
Everyone knows Friday the 13th is considered an unlucky day. But why does it have such a bad reputation ? One reason is that both Friday and the number 13 have some troubled ties to Christianity . <1>Ever since , the day has been connected with “ bad luck ” . In the Middle Ages , for instance , weddings were not held on Fridays . <2>Friday was also unlucky in medieval times because it was “ hangman ’s day ” . As for the number 13 , seating 13 people at a table was seen as bad luck because Judas Iscariot , the disciple who betrayed Jesus , is said to have been the 13th guest at The Last Supper . It ’s unclear exactly when Friday and the number 13 became connected . <3>In 1907 , Thomas Lawson wrote a book titled Friday , the Thirteenth, which described a stockbroker choosing this day to bring down Wall Street. Vyse , who specializes in the psychology of superstitions , says this is a kind of belief . <4>Although there ’s evidence that believing in lucky symbols is helpful , Friday the 13th represents a kind of fear . In fact , fear of Friday the 13th has a name : paraskavedekatriaphobia . In Vyse ’s view , ” <5>
#candidates:
[’We only know there are no mentions of Friday the 13th before the 19th century .’,
’In fact , in Italy,13 is generally considered a lucky number .’,
’It ’s a way for people to “control the uncontrollable” and manage the anxiety that comes with uncertain situations.’,
’It was on a Friday that Jesus was put to death .’,
’Fridays are regarded as an unlucky day and thirteen as an unlucky number.’,
’We would be better off if no one had ever taught them to us.’,
’Likewise , it was not a day someone would set out on a journey.’]
#Output:
{
"output": [
{"position": 1,
"sentence": "It was on a Friday that Jesus was put to death ."},
{"position": 2,
"sentence": "Likewise , it was not a day someone would set out on a journey ."},
{"position": 3,
"sentence": "We only know there are no mentions of Friday the 13th before the 19th century ."},
{"position": 4,
"sentence": "It ’s a way for people to “ control the uncontrollable ” and manage the anxiety that comes with uncertain situations ."},
{"position": 5,
"sentence": "We would be better off if no one had ever taught them to us ."}]
}
 	
#passage:
##passage##
#candidates:
##candidates##
#output:

Table 25: Prompt for connecting skill evaluation.

D.4 Organizing

Prompt
Given a passage and a list of subheadings derived from that passage, your task is to divide the passage into sections, ensuring that each section corresponds to a subheading in the list. You should insert each subheading into the appropriate position in the passage to achieve text segmentation. Output the subheading and its position in turn in JSON format. The key of the JSON should be "output" and its corresponding value should be a list. Response with the JSON only! Here is an example: #Passage: (1) Based on current temperature forecasts, a record-breaking148 million Americansare expected to experience CSI values of 3 or higher on August 2nd. (2) That means their local temperatures would be made at least 3 times more likely because of climate change. (3) Hazardous heat conditions are projected from theRocky Mountainsto thesoutheastern and Mid-Atlantic United States, with late-week and weekend high temperatures reaching themid-to-upper 90sandlow 100s (°F). (4) The National Weather Service forecasts several days of Major to Extreme Heat Risk, from theCentral Plainsto thesoutheastern United States, increasing the likelihood of health impacts for individuals without proper hydration and cooling. (5) High humidity combined with excessive temperatures will lead to dangerous heat index values. (6) Feels-like conditions exceeding110°Fare forecasted across the GreatPlains, Mississippi Valley,andsoutheastern United States. (7) The significant heat dome responsible for these high temperatures is expected to expand across a majority of the Lower 48 statesthrough the first week of August. #Subheadings: [’Climate Shift Index exposure approaches record’, ’How unusual is the forecasted heat?’] #Output: {"Output": [{"subheading": ’Climate Shift Index exposure approaches record’, "position": 1}, {"subheading": ’How unusual is the forecasted heat?’, "position": 3}] } #Passage: ##passage## #Subheadings: ##subheadings## #Output:

Table 26: Prompt for organizing skill evaluation.

D.5 Selecting

Prompt
Given a passage, you need to select the most important sentences that can serve as a summary of the passage. Sentences in the passage are separated by <S>. Output these selected sentences in a list format, organized in descending order of importance. Response with the List only! Here is an example: #Passage: 13 March 2012 Last updated at 18:31 GMT<S>Nan Weidong and Nan Weiping have been transforming vegetables into musical instruments for two years.<S>Their dad was a music teacher and encouraged them to be musical from a young age - but carrot panpipes probably weren’t what he had in mind!<S>Weidong says it’s important the veg is fresh - otherwise it risks being out of tune.<S>And no vegetable is too much of a challenge: they’ve turned a sweet potato into an ocarina, a bamboo shoot has become a flute, and a yam has doubled up as a whistle.<S>Watch the clip to see them in action! #Output: {"Output": ["Nan Weidong and Nan Weiping have been transforming vegetables into musical instruments for two years.", "Their dad was a music teacher and encouraged them to be musical from a young age - but carrot panpipes probably weren’t what he had in mind!", "And no vegetable is too much of a challenge: they’ve turned a sweet potato into an ocarina, a bamboo shoot has become a flute, and a yam has doubled up as a whistle."] } #Passage: ##passage## #Output:

Table 27: Prompt for selecting skill evaluation.