Make Your LLM Fully Utilize the Context

Shengnan An, Zexiong Ma, Zeqi Lin,
Nanning Zheng, Jian-Guang Lou, Weizhu Chen

西安交通大學
XI'AN JIAOTONG UNIVERSITY

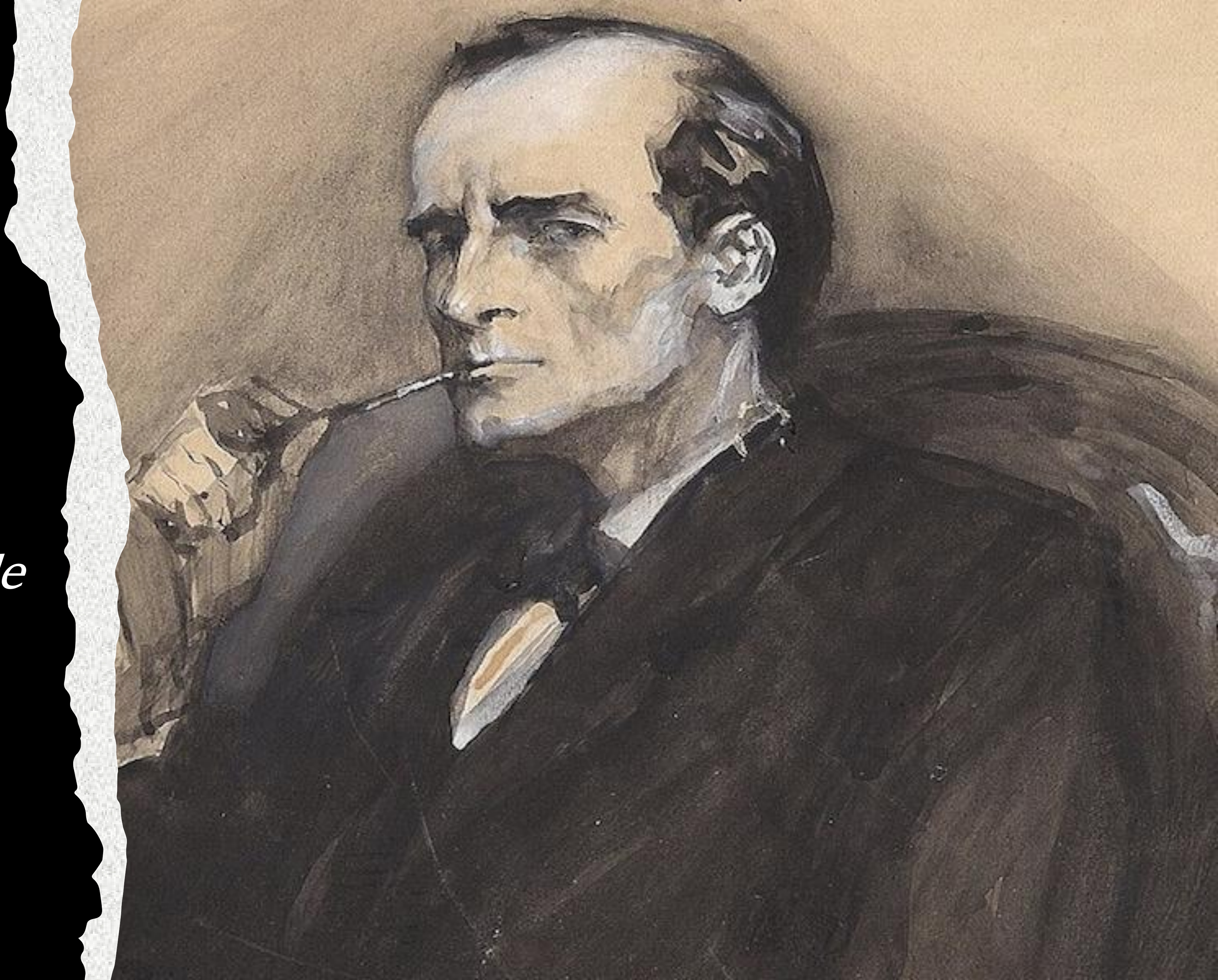
Microsoft

北京大學
PEKING UNIVERSITY

To A Great Mind,

Nothing Is Little.

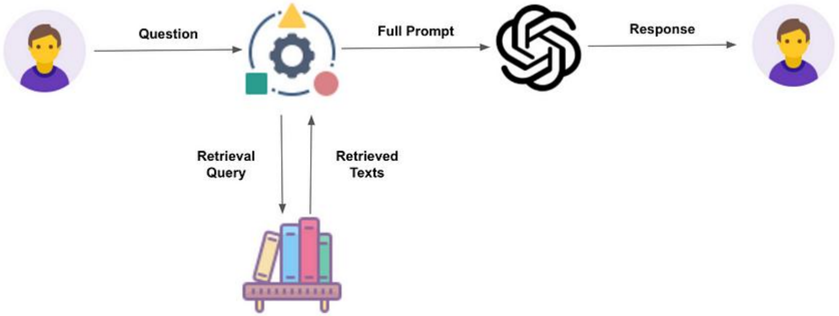
---Arthur Conan Doyle



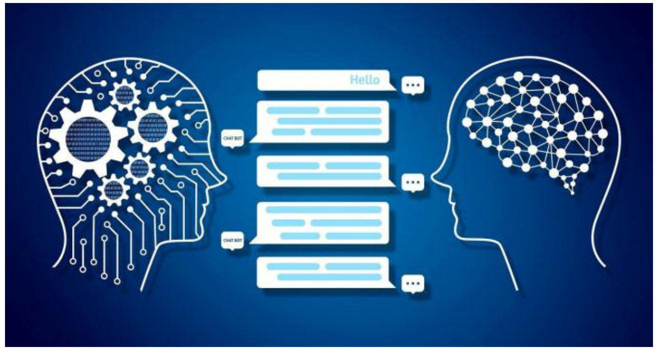
Long-Context LLM

Long-context large language model is currently receiving extensive attention from both academia and industry. The training context windows of many contemporary LLMs have been expanded to tens of thousands of tokens, thereby enabling these models to process extensive context as input.

Retrieval Augmented Generation (RAG)



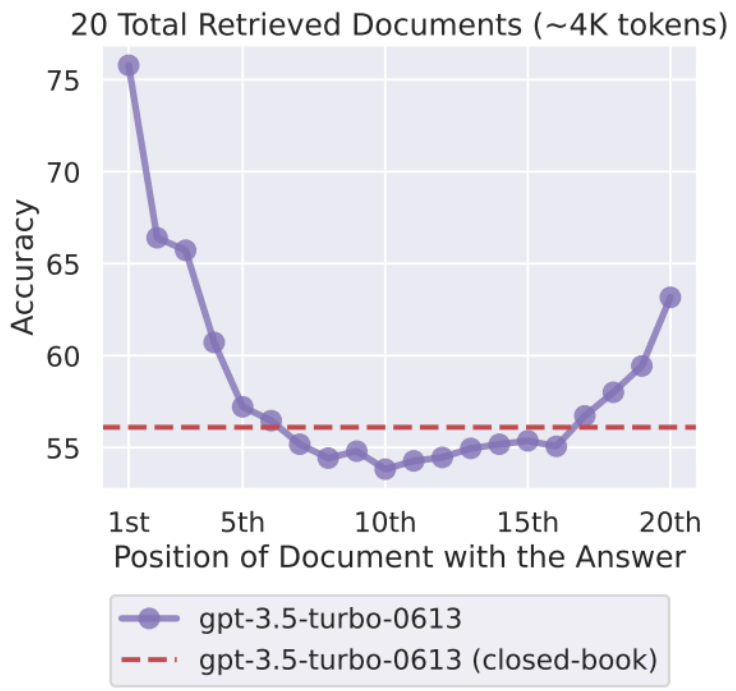
Life-Long Conversations



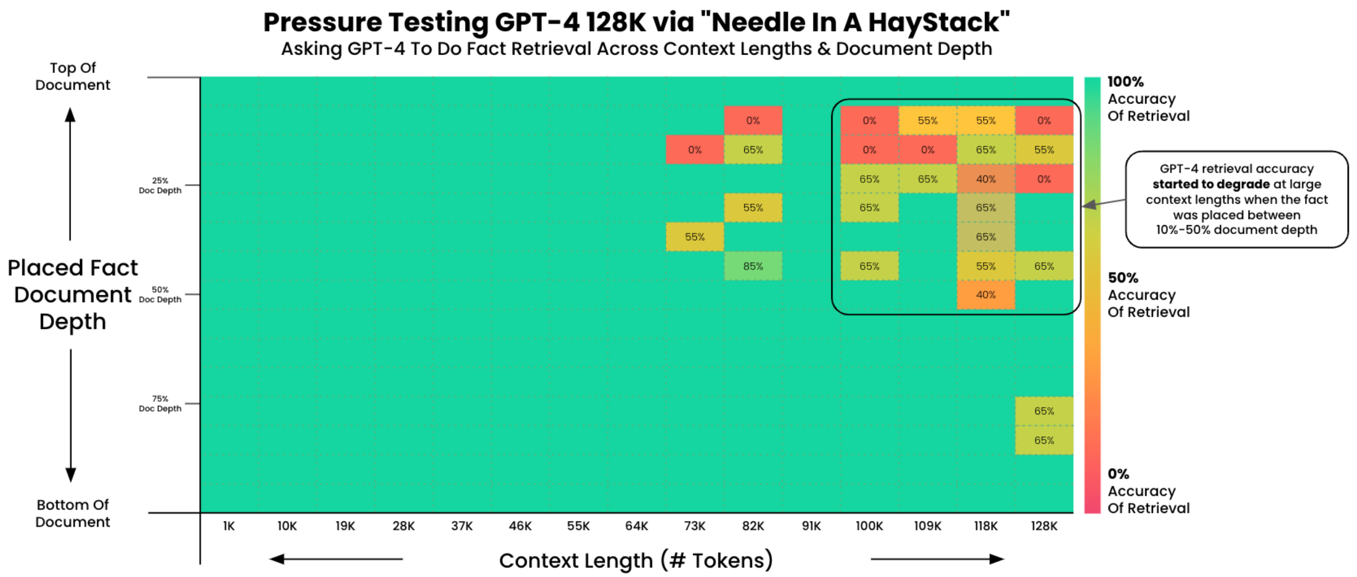
Long Document QA & Summarization

Challenge: Lost in the Middle

Contemporary long-context LLMs struggle to effectively and robustly utilize all the information provided in the context, known as the lost-in-the-middle challenge [1, 2]. It implies that while the LLM can comprehend the information at the head and tail of the long context, it often overlooks the information in the middle.



20 Total Retrieved Documents (~4K tokens)



Pressure Testing GPT-4 128K via "Needle In A HayStack"

Asking GPT-4 To Do Fact Retrieval Across Context Lengths & Document Depth

GPT-4 retrieval accuracy **started to degrade** at large context lengths when the fact was placed between 10%-50% document depth

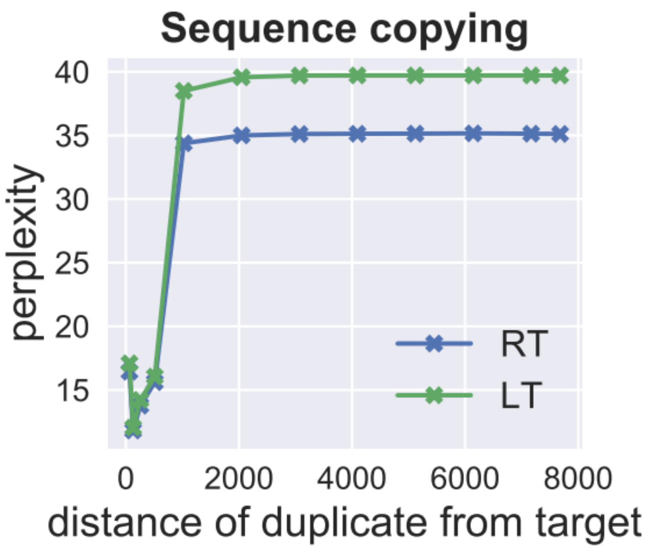
Goal: Test GPT-4 Ability To Retrieve Information From Large Context Windows
A fact was placed within a document. GPT-4 (1106-preview) was then asked to retrieve it. The output was evaluated for accuracy.
This test was run at 15 different document depths (top > bottom) and 15 different context lengths (1K >128K tokens).
2x tests were run for larger contexts for a larger sample size.

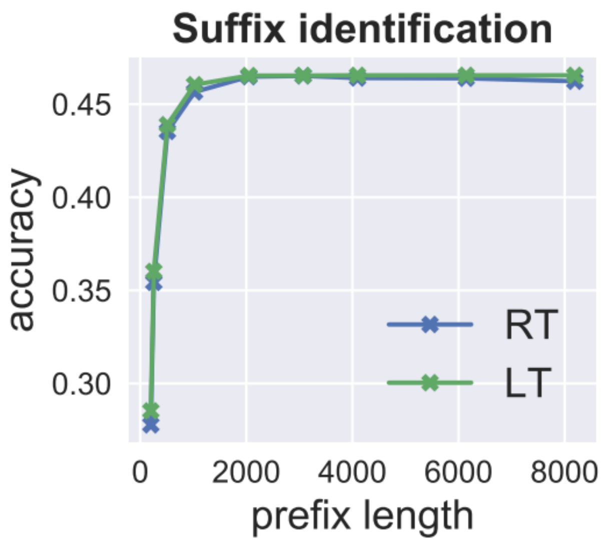
[1] Nelson F. Liu, et al. "Lost in the middle: How language models use long contexts." TACL 2024.
[2] Peng Xu, et al. "Retrieval meets Long Context Large Language Models." ICLR 2023.

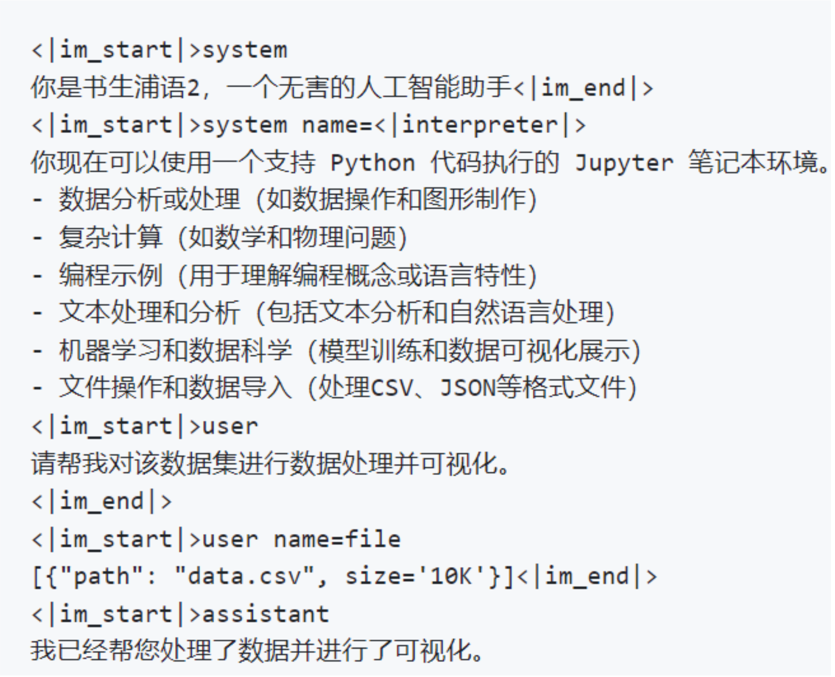
Challenge: Lost in the Middle

We hypothesize that the root cause of lost-in-the-middle stems from the unintentional bias hidden in the general training data.

- In auto-regressive pre-training, the loss on predicting the next token is more likely to be influenced by a few nearby pre-tokens rather than long-distance tokens [1].
- For supervised fine-tuning and alignment, the system message, which strongly influences the generation of the response, is typically presented at the beginning of the context [2].







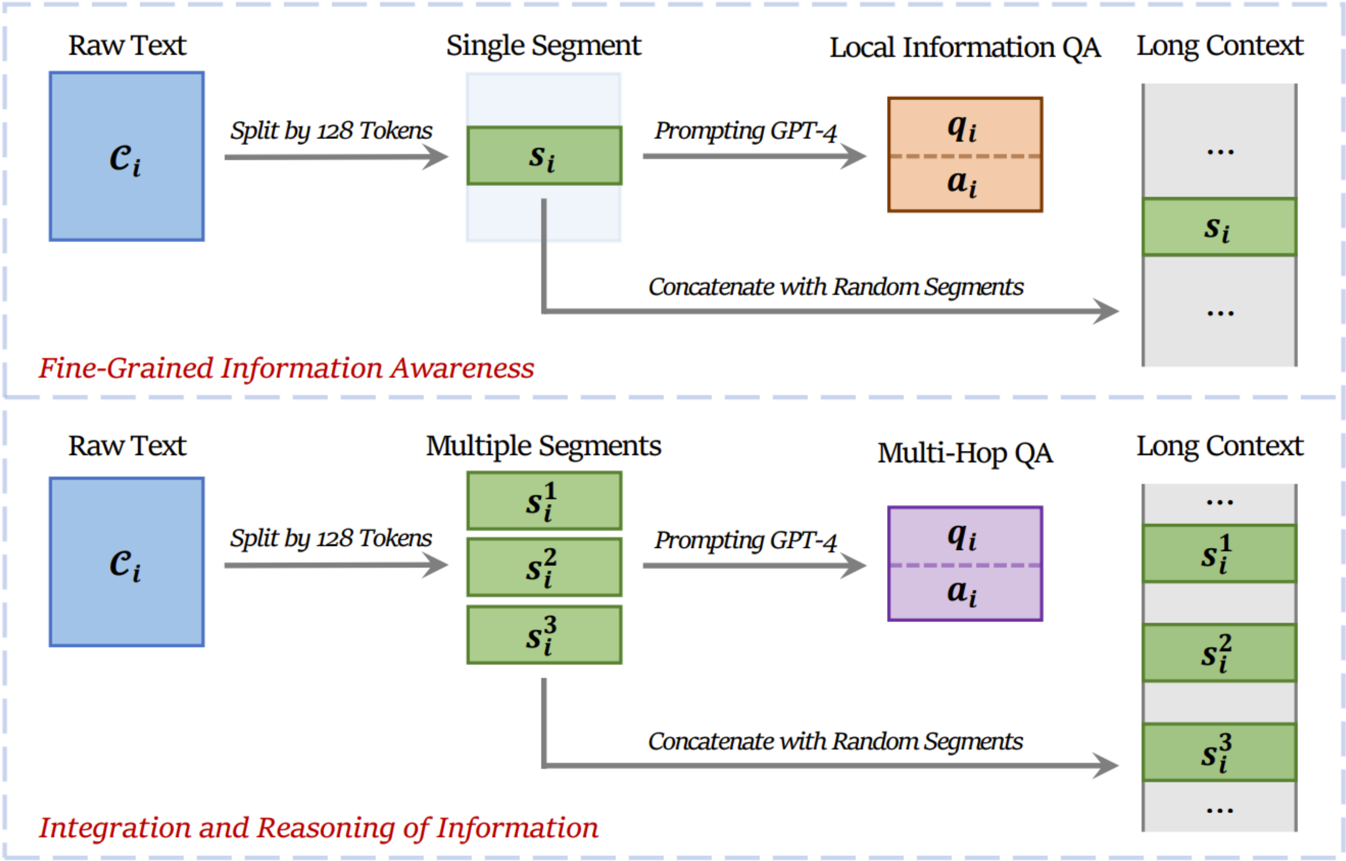
```
<|im_start|>system
你是书生浦语2，一个无害的人工智能助手<|im_end|>
<|im_start|>system name=<|interpreter|>
你现在可以使用一个支持 Python 代码执行的 Jupyter 笔记本环境。
- 数据分析或处理（如数据操作和图形制作）
- 复杂计算（如数学和物理问题）
- 编程示例（用于理解编程概念或语言特性）
- 文本处理和分析（包括文本分析和自然语言处理）
- 机器学习和数据科学（模型训练和数据可视化展示）
- 文件操作和数据导入（处理CSV、JSON等格式文件）
<|im_start|>user
请帮我对该数据集进行数据处理并可视化。
<|im_end|>
<|im_start|>user name=file
[{"path": "data.csv", size='10K'}]<|im_end|>
<|im_start|>assistant
我已经帮您处理了数据并进行了可视化。
```

[1] Simeng Sun, et al. "Do Long-Range Language Models Actually Use Long-Range Context?" EMNLP 2021.

[2] InternLM2 Technical Report.

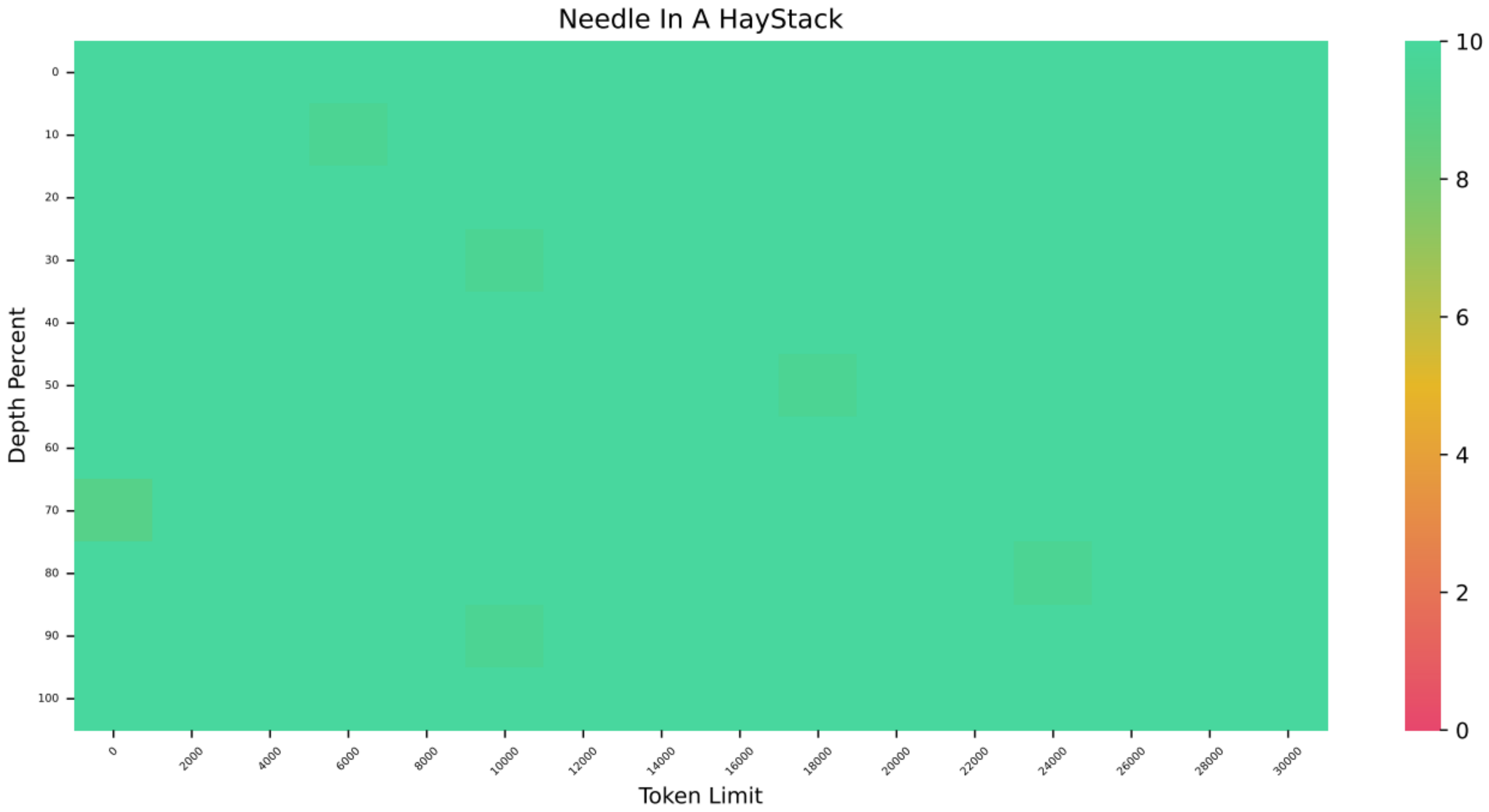
Solution: IN2 Training

To overcome lost-in-the-middle, our work introduces Information-Intensive training, a purely data-driven solution to explicitly teach the model that the crucial information can be intensively present throughout the context.



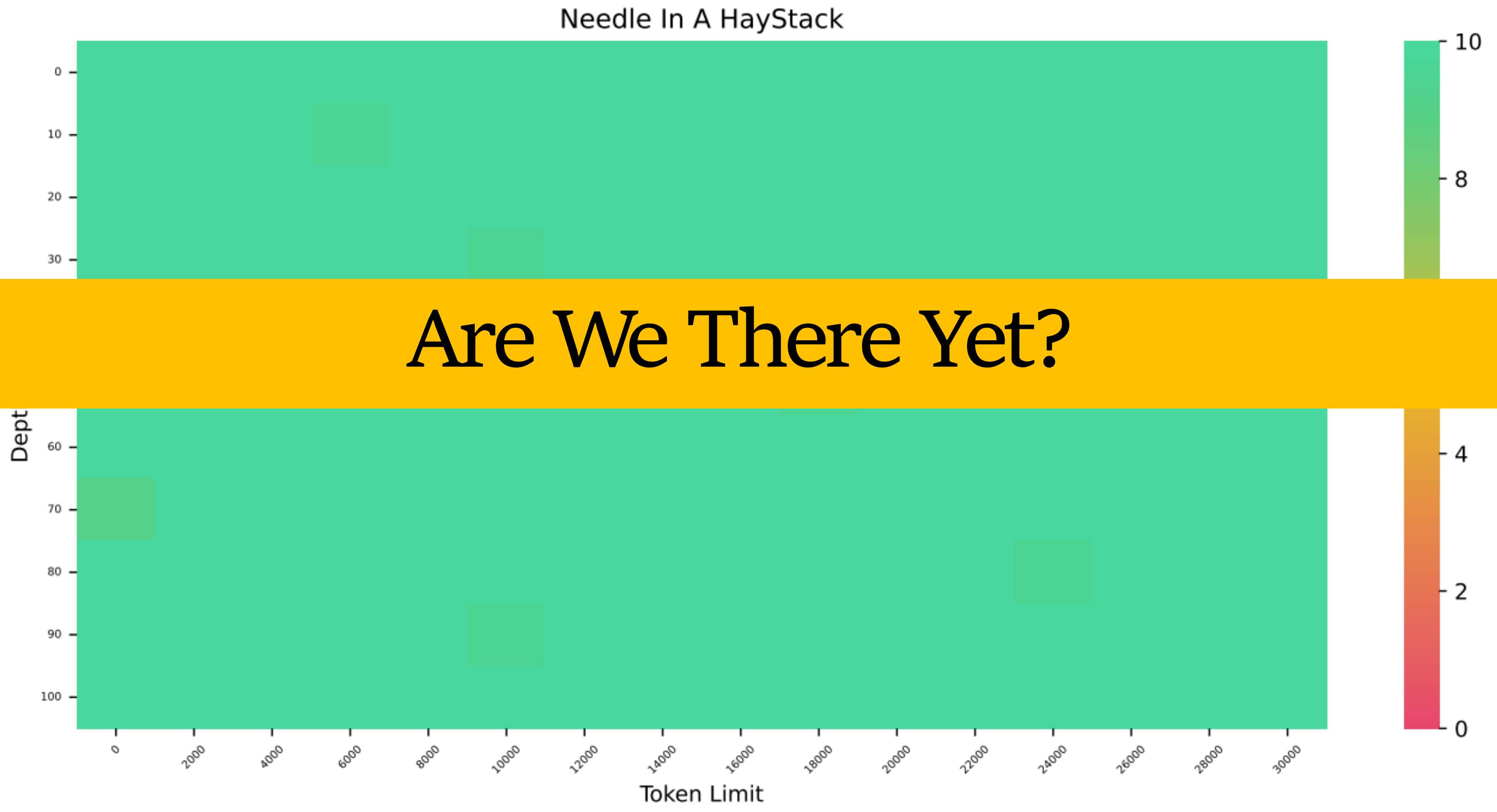
Model: FILM-7B

Through applying IN2 Training on Mistral-7B, we present FILM-7B (Fill-in-the-Middle) with a 32K context window. FILM-7B achieves near-perfect performance on Needle-in-the-Haystack.



Model: FILM-7B

Through applying IN2 Training on Mistral-7B, we present FILM-7B (Fill-in-the-Middle) with a 32K context window. FILM-7B achieves near-perfect performance on Needle-in-the-Haystack.



Are We There Yet?

Model: FILM-7B

Through applying IN2 Training on Mistral-7B, we present FILM-7B (Fill-in-the-Middle) with a 32K context window. FILM-7B achieves near-perfect performance on Needle-in-the-Haystack.

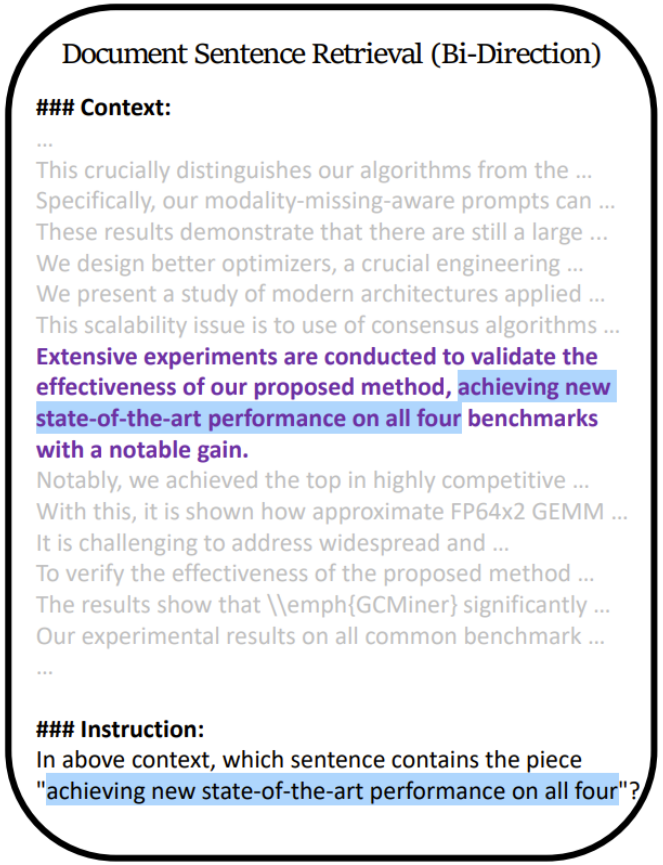
Needle-in-the-Haystack is not enough for long-context probing:

- It employs a document-style context, which LLMs could be quite familiar with due to the pre-training on natural language corpora.
- The forward retrieval pattern in Needle-in-the-Haystack may simplify the difficulty of information seeking in the long context, due to the mechanism of inductive head.

> *Example in Needle-in-the-Haystack*
>
> **Context:**
> … … The best thing to do in San Francisco is eat a sandwich and sit in Dolores Park on a sunny day … …
>
> **Question:**
> What is the best thing to do in San Francisco?

VAL Probing

We introduce Various Long-context (VAL) Probing which encompasses three context styles (document, code, and structured-data context) and three information retrieval patterns (forward, backward, and bi-directional retrieval).



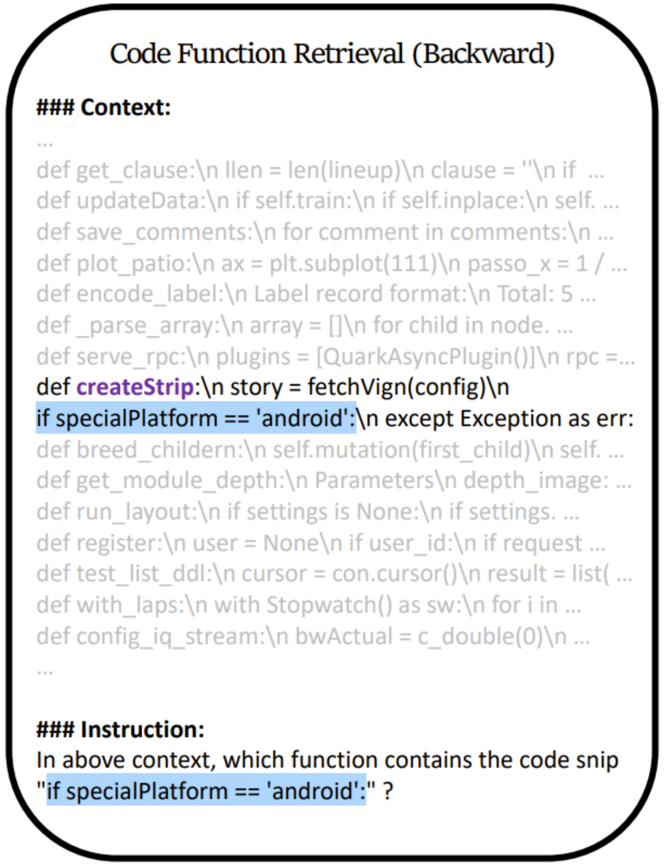
Document Sentence Retrieval (Bi-Direction)

Context:

...
This crucially distinguishes our algorithms from the ...
Specifically, our modality-missing-aware prompts can ...
These results demonstrate that there are still a large ...
We design better optimizers, a crucial engineering ...
We present a study of modern architectures applied ...
This scalability issue is to use of consensus algorithms ...
Extensive experiments are conducted to validate the effectiveness of our proposed method, achieving new state-of-the-art performance on all four benchmarks with a notable gain.
Notably, we achieved the top in highly competitive ...
With this, it is shown how approximate FP64x2 GEMM ...
It is challenging to address widespread and ...
To verify the effectiveness of the proposed method ...
The results show that \\emph{GCMiner} significantly ...
Our experimental results on all common benchmark ...
...

Instruction:

In above context, which sentence contains the piece "achieving new state-of-the-art performance on all four"?



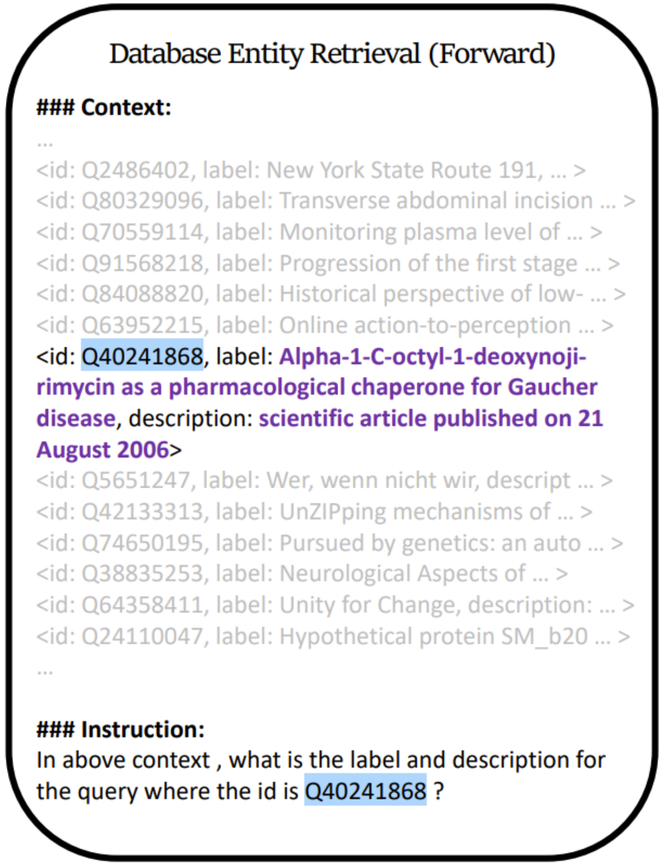
Code Function Retrieval (Backward)

Context:

```
...
def get_clause:\n llen = len(lineup)\n clause = ''\n if  ...
def updateData:\n if self.train:\n if self.inplace:\n self. ...
def save_comments:\n for comment in comments:\n ...
def plot_patio:\n ax = plt.subplot(111)\n passo_x = 1 / ...
def encode_label:\n Label record format:\n Total: 5 ...
def _parse_array:\n array = []\n for child in node. ...
def serve_rpc:\n plugins = [QuarkAsyncPlugin()]\n rpc =...
def createStrip:\n story = fetchVign(config)\n
if specialPlatform == 'android':\n except Exception as err:
def breed_childern:\n self.mutation(first_child)\n self. ...
def get_module_depth:\n Parameters\n depth_image: ...
def run_layout:\n if settings is None:\n if settings. ...
def register:\n user = None\n if user_id:\n if request ...
def test_list_ddl:\n cursor = con.cursor()\n result = list( ...
def with_laps:\n with Stopwatch() as sw:\n for i in ...
def config_iq_stream:\n bwActual = c_double(0)\n ...
...
```

Instruction:

In above context, which function contains the code snip "if specialPlatform == 'android':" ?



Database Entity Retrieval (Forward)

Context:

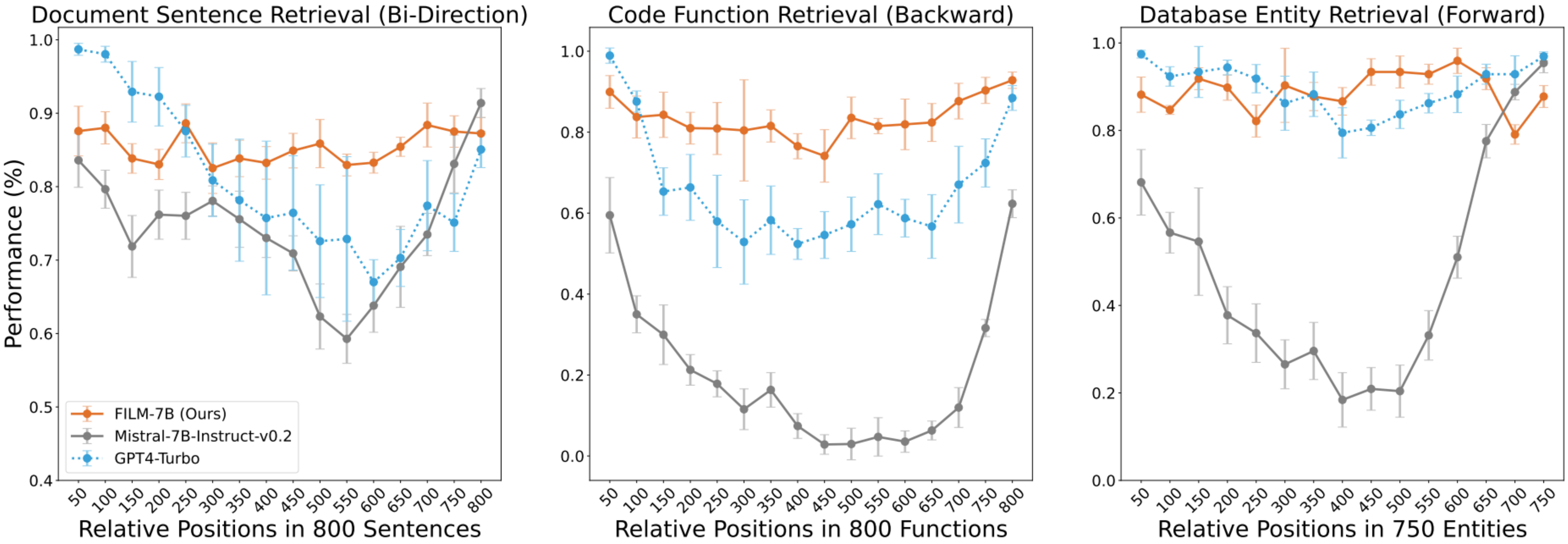
...
<id: Q2486402, label: New York State Route 191, ... >
<id: Q80329096, label: Transverse abdominal incision ... >
<id: Q70559114, label: Monitoring plasma level of ... >
<id: Q91568218, label: Progression of the first stage ... >
<id: Q84088820, label: Historical perspective of low- ... >
<id: Q63952215, label: Online action-to-perception ... >
<id: Q40241868, label: **Alpha-1-C-octyl-1-deoxynojirimycin as a pharmacological chaperone for Gaucher disease**, description: **scientific article published on 21 August 2006**>
<id: Q5651247, label: Wer, wenn nicht wir, descript ... >
<id: Q42133313, label: UnZIPping mechanisms of ... >
<id: Q74650195, label: Pursued by genetics: an auto ... >
<id: Q38835253, label: Neurological Aspects of ... >
<id: Q64358411, label: Unity for Change, description: ... >
<id: Q24110047, label: Hypothetical protein SM_b20 ... >
...

Instruction:

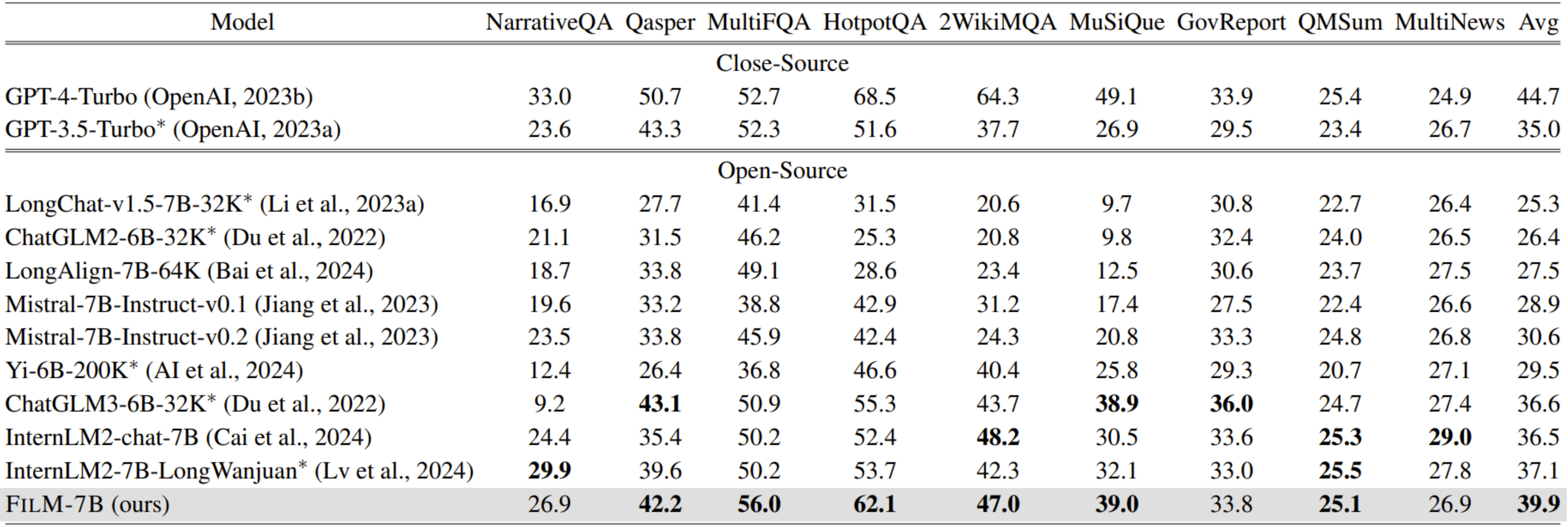
In above context , what is the label and description for the query where the id is Q40241868 ?

Performance on VAL Probing

The results on VAL Probing demonstrate that our Information-Intensive training can significantly mitigates the lost-in-the-middle problem.

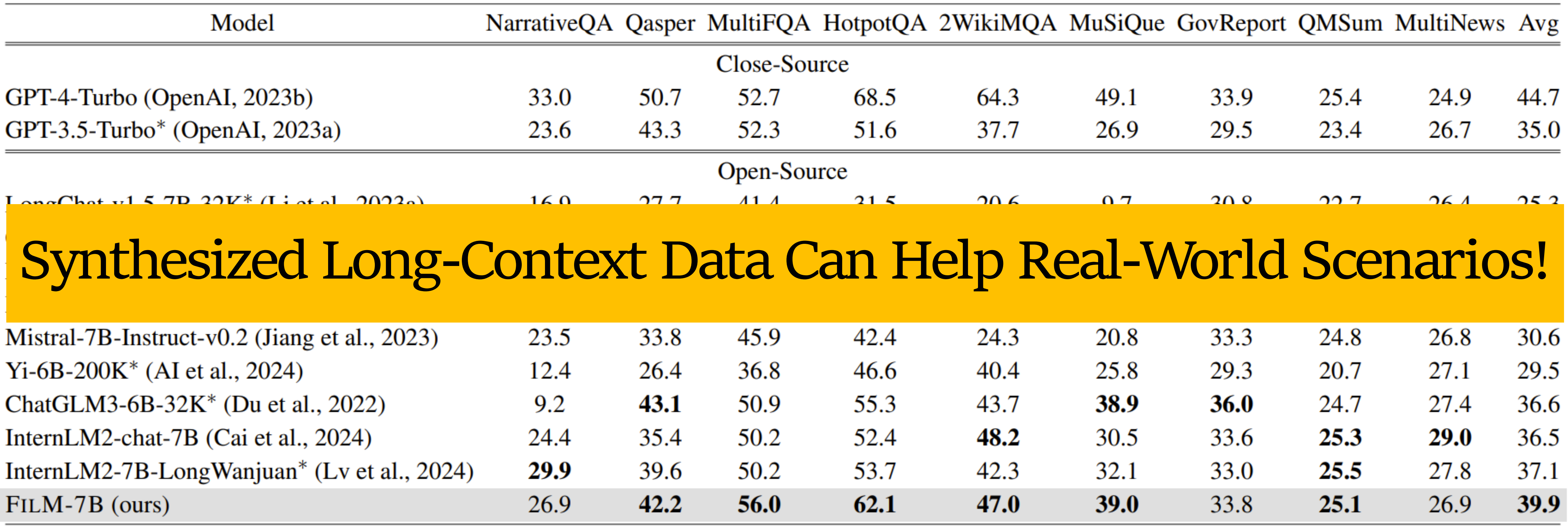


Real-World Long-Context Tasks



Model	NarrativeQA	Qasper	MultiFQA	HotpotQA	2WikiMQA	MuSiQue	GovReport	QMSum	MultiNews	Avg
					Close-Source					
GPT-4-Turbo (OpenAI, 2023b)	33.0	50.7	52.7	68.5	64.3	49.1	33.9	25.4	24.9	44.7
GPT-3.5-Turbo* (OpenAI, 2023a)	23.6	43.3	52.3	51.6	37.7	26.9	29.5	23.4	26.7	35.0
					Open-Source					
LongChat-v1.5-7B-32K* (Li et al., 2023a)	16.9	27.7	41.4	31.5	20.6	9.7	30.8	22.7	26.4	25.3
ChatGLM2-6B-32K* (Du et al., 2022)	21.1	31.5	46.2	25.3	20.8	9.8	32.4	24.0	26.5	26.4
LongAlign-7B-64K (Bai et al., 2024)	18.7	33.8	49.1	28.6	23.4	12.5	30.6	23.7	27.5	27.5
Mistral-7B-Instruct-v0.1 (Jiang et al., 2023)	19.6	33.2	38.8	42.9	31.2	17.4	27.5	22.4	26.6	28.9
Mistral-7B-Instruct-v0.2 (Jiang et al., 2023)	23.5	33.8	45.9	42.4	24.3	20.8	33.3	24.8	26.8	30.6
Yi-6B-200K* (AI et al., 2024)	12.4	26.4	36.8	46.6	40.4	25.8	29.3	20.7	27.1	29.5
ChatGLM3-6B-32K* (Du et al., 2022)	9.2	**43.1**	50.9	55.3	43.7	**38.9**	**36.0**	24.7	27.4	36.6
InternLM2-chat-7B (Cai et al., 2024)	24.4	35.4	50.2	52.4	**48.2**	30.5	33.6	**25.3**	**29.0**	36.5
InternLM2-7B-LongWanjuan* (Lv et al., 2024)	**29.9**	39.6	50.2	53.7	42.3	32.1	33.0	**25.5**	27.8	37.1
FILM-7B (ours)	26.9	**42.2**	**56.0**	**62.1**	**47.0**	**39.0**	33.8	**25.1**	26.9	**39.9**

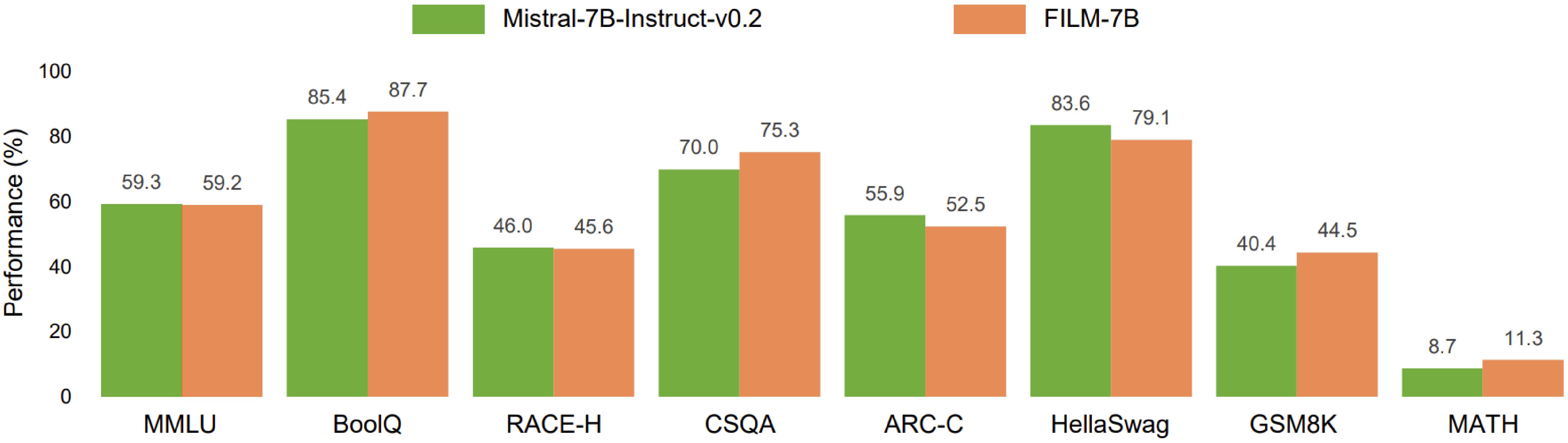
Real-World Long-Context Tasks



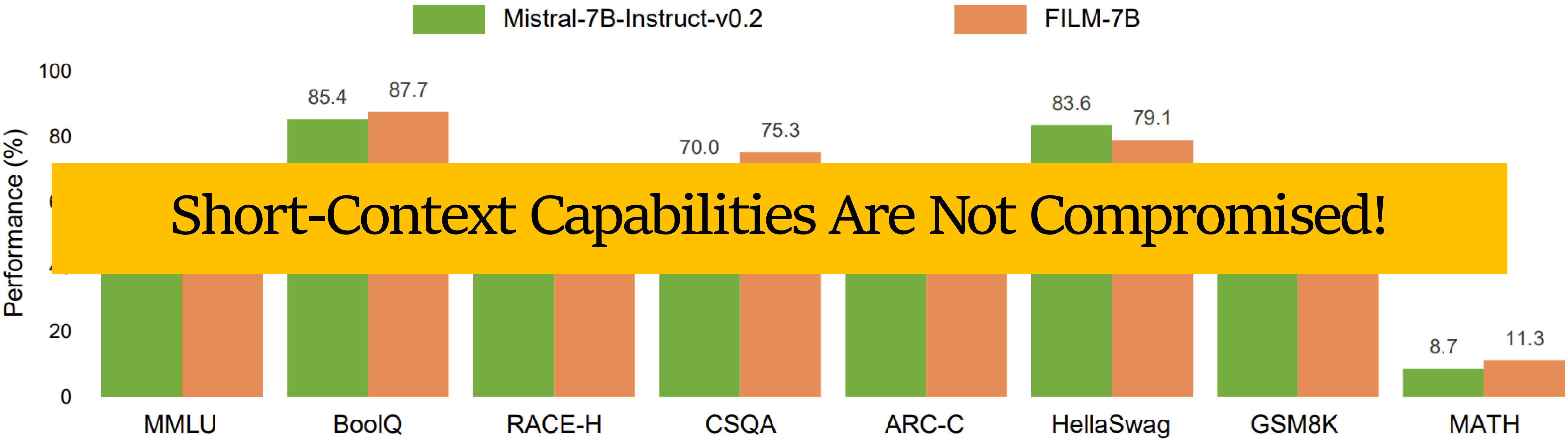
Model	NarrativeQA	Qasper	MultiFQA	HotpotQA	2WikiMQA	MuSiQue	GovReport	QMSum	MultiNews	Avg
					Close-Source					
GPT-4-Turbo (OpenAI, 2023b)	33.0	50.7	52.7	68.5	64.3	49.1	33.9	25.4	24.9	44.7
GPT-3.5-Turbo* (OpenAI, 2023a)	23.6	43.3	52.3	51.6	37.7	26.9	29.5	23.4	26.7	35.0
					Open-Source					
LongChat-v1.5-7B-32K* (Li et al., 2023a)	16.9	27.7	41.4	31.5	20.6	9.7	30.8	22.7	26.4	25.3
Mistral-7B-Instruct-v0.2 (Jiang et al., 2023)	23.5	33.8	45.9	42.4	24.3	20.8	33.3	24.8	26.8	30.6
Yi-6B-200K* (AI et al., 2024)	12.4	26.4	36.8	46.6	40.4	25.8	29.3	20.7	27.1	29.5
ChatGLM3-6B-32K* (Du et al., 2022)	9.2	**43.1**	50.9	55.3	43.7	**38.9**	**36.0**	24.7	27.4	36.6
InternLM2-chat-7B (Cai et al., 2024)	24.4	35.4	50.2	52.4	**48.2**	30.5	33.6	**25.3**	**29.0**	36.5
InternLM2-7B-LongWanjuan* (Lv et al., 2024)	**29.9**	39.6	50.2	53.7	42.3	32.1	33.0	**25.5**	27.8	37.1
FILM-7B (ours)	26.9	**42.2**	**56.0**	**62.1**	**47.0**	**39.0**	33.8	**25.1**	26.9	**39.9**

Synthesized Long-Context Data Can Help Real-World Scenarios!

Short-Context Tasks



Short-Context Tasks



Short-Context Capabilities Are Not Compromised!

Take-Away

1. We present IN2 Training to overcome lost-in-the-middle.
2. We introduce VAL Probing to comprehensively evaluate long-context information utilization.
3. We release FILM-7B, a powerful 32K-context model.

Github Link: https://github.com/microsoft/FILM