Supplementary Material for ECIR 2025 Submission

Prevent Jargon-Induced Hallucination with Reflective Retrieval Augmented Generation

1 Fine-tuning or Retrieval Augmented Generation?

Knowledge injection via fine-tuning has several significant drawbacks. For instance, when fine-tuned on a knowledge statement like "A is B," the fine-tuned LLM can correctly answer "What is A?" but fails to answer "What is B?" with "A" for arbitrary A and B. This phenomenon is famously known as The Reversal Curse (Berglund et al., 2023). Although remedies such as generating reversed training data (Golovneva et al., 2024) have been proposed, they require higher training costs and do not guarantee that the tuned LLM will answer all possible forms of a query. Additionally, incorporating knowledge through fine-tuning necessitates a new fine-tuning job for each new piece of knowledge, which incurs computational costs and hinders efficient integration of new information. The amount of knowledge a model can effectively incorporate depends on the capacity of the fine-tuned model part (Roberts et al., 2020), while excessive fine-tuning may lead to catastrophic forgetting, where the model forgets previously learned knowledge (Zhai et al., 2024).

In contrast, RAG does not suffer from these drawbacks. The Reversal Curse, observed in fine-tuning methods, does not occur when knowledge statements are presented in-context, as part of the prompt. In RAG, the LLM learns knowledge statements in-context, significantly improving its reasoning capacity and enabling efficient instruction prompt tuning (Singhal et al., 2023). Furthermore, RAG does not require model retraining and can efficiently incorporate new knowledge corpora. These properties make RAG a superior choice for industrial knowledge bases.

2 Question-Answering Data Examples

Here we show a few non-confidential instances of the evaluation data used in the question-answering experiment, as follows:

Who decides ACT timing SPECS? a. Memory Team. b. System Team. c. JEDEC/ONFI d. Customer. **Answer:** c

In any of the 2D NAND dies, CMD/ADDR protocol is of what nature? 1. Legacy 2. DDR1 3. DDR2 4. Depends on that particular project/technology node. **Answer:** 1

In Dynamic timing analysis, timing parameters margin is checked without doing any simulations. a. True b. False. **Answer:** b

What is the VCCQ level for DDR3 standard? 1. 1.8V 2. 1.7V 3. 1.6V 4. 1.5V 5. 1.2V **Answer:** 4

What is the state of ALEx,CLEx during DATA IN operation? a. 00 b. 11 c. 10 d. Don't care. **Answer:** a
We need 5ns clock period to achieve 400MBps in DDR. To achieve 50MBps in SDR, what should be the clock period? 1. 10ns 2. 20ns 3. 40ns 4. 80ns **Answer:** 2
Why do we use dummy transistors? a. protect the actual transistors while fabrication b. Can be used as spare transistors to be used in refinement of the circuit. c. To create uniform environment for pair transistors. d. All of the above **Answer:** d
Which of the following tasks is not the responsibility of ACT team? 1. IO design 2. Data path design 3. Pad order design 4. Package design **Answer:** 4
Which of the following factors affect Electromigration in the circuit? 1. Number of contacts/vias at the connecting junction of two metal layers 2. Current density in the metal layer 3. Temperature 4. All of above **Answer:** 4
What parameter Receiver skew affects largely in the design? a) Input Voltage b) Duty Cycle c) Input Slew Rate d) Data reception **Answer:** b

3 Abbreviation Identification Experiment

3.1 Synthetic Dataset Generation Template

Below is the question template and the list of random abbreviations used for generating random abbreviations in the abbreviation identification experiment.

```
question_templates = [
    # Templates with one abbreviation
    "What does the abbreviation {abbr1} stand for?",
    "Can you explain the meaning of {abbr1}?",
    "What is the full form of {abbr1}?",
    "{abbr1} is an abbreviation for what?",

    # Templates with two abbreviations
    "What do the abbreviations {abbr1} and {abbr2} mean?",
    "In the case where {abbr1} > 0.5, how much should {abbr2} be?",
    "What is the relationship between {abbr1} and {abbr2}?",

    # Templates with three abbreviations
    "Consider {abbr1} = 1.5 and {abbr2} < 0.1, what would {abbr3} be?",
    "{abbr1} and {abbr2} are the same. Should {abbr3} be high or low?",
    "What is the state of {abbr1}, {abbr2} during {abbr3} operation?",

    # Templates with four abbreviations
    "We need 10ns {abbr1} to achieve 40{abbr2} in {abbr3}. What should be {abbr4}?",
    "In any of the {abbr1}, {abbr2}/{abbr3}/{abbr4} should be what nature?",

    # Template with five abbreviations
    "{abbr1}=10, {abbr2}=5, {abbr3}<0.1 in {abbr4}. How should I set {abbr5}?"
]

random_abbreviation_list = ['TS', 'IE', 'MI', 'SF', 'MP', 'UM', 'ES', 'PE', 'UW', '
    SU', 'FSU', 'QMB', 'KPU', 'VMT', 'ESO', 'ARI', 'SPA', 'MTD', 'GTC', 'ODV', 'SLBG
    ', 'MUBO', 'ROSN', 'VPPL', 'PIOF', 'CCPP', 'MBST', 'UTUU', 'NIRE', 'STUP']
```

3.2 Sample Experiment Results and Fail Cases

Here we present the sample experiment results from the abbreviation identification experiment. We emphasize on demonstrating the failure modes, which is different across LLMs. In the next subsection, we will discuss the failure modes.

3.2.1 Meta Llama 3

```
[
    {
        "Question": "What is the full form of KPU?",
        "Response": "[\"KPU\"]" # correct
    },
    {
        "Question": "What is the full form of ARI?",
        "Response": "[\"ARI\"]" # correct
    },
    {
        "Question": "Can you explain the meaning of GTC?",
        "Response": "[\"GTC\", \"Good Till Cancelled\"]" # incorrect
    },
    {
        "Question": "Can you explain the meaning of SPA?",
        "Response": "[\"Single-Page Application\", \"SPA\"]" # incorrect
    },
    {
        "Question": "In any of the UTUU, ES/NIRE/MUBO should be what nature?",
        "Response": "[\"UTUU\", \"ES/NIRE/MUBO\"]" # incorrect
    },
]
```

3.2.2 Mistral

```
[
    {
        "Question": "We need 10ns ARI to achieve 40MI in MUBO. What should be PIOF?"
            ,
        "Response": "[\"ARI\", \"MI\", \"MUBO\", \"PIOF\"]" # correct
    },
    {
        "Question": "We need 10ns NIRE to achieve 40MP in STUP. What should be IE?",
        "Response": "[\"10ns NIRE\", \"40MP in STUP\", \"IE\"]" #incorrect
    },
    {
        "Question": "We need 10ns MBST to achieve 40ROSN in SPA. What should be UW?"
            ,
        "Response": "[\"10ns MBST\", \"40ROSN\", \"SPA\", \"UW\"]" # incorrect
    },
    {
        "Question": "In any of the UTUU, ES/NIRE/MUBO should be what nature?",
        "Response": "[\"UTUU\", \"ES/NIRE/MUBO\"]" # incorrect
    },
]
```

3.2.3 Shisa

```
[
    {
        "Question": "PIOF is an abbreviation for what?",
        "Response": "[\"PIOF\"]</s>" # correct
    },
    {
        "Question": "What is the full form of IE?",
        "Response": "[\"IE\", \"Internet Explorer\"]\n\nPlease let me know if you
            need anything else.</s>" # incorrect
    },
    {
        "Question": "SPA is an abbreviation for what?",
        "Response": "[\"SPA\"]\n\nPlease let me know if I can assist you further.</s
            >" # incorrect
    },
    {
        "Question": "Can you explain the meaning of SPA?",
        "Response": "[\"SPA\"]\n\nThe text translated to Japanese is:\n[\u300cSPA\
            u300d]</s>" # incorrect
    },
    {
        "Question": "Can you explain the meaning of SU?",
        "Response": "[\"SU\"]\n\nThe text translated to Japanese is:\n### \u6307\
            u793a:\n\u300cSU\u300d\u3068\u3044\u3046\u610f\u5473\u3092\u8aac\u660e\
            u3067\u304d\u307e\u3059\u304b\uff1f\u3053\u306e\u8cea\u554f\u304b\u3089\
            u3001\u5c02\u9580\u7528\u8a9e\u3084\u7565\u8a9e\u3092\u7279\u5b9a\u3057\
            u3066\u304f\u3060\u3055\u3044\u3002\u6b21\u306e\u5f62\u5f0f\u3067\u30ea\
            u30b9\u30c8\u30a2\u30c3\u30d7\u3057\u3066\u304f\u3060\u3055\u3044: [\
            u300c\u5c02\u9580\u7528\u8a9e1\u300d\u3001\u300c\u5c02\u9580\u7528\
            u8a9e2\u300d\u3001...]\n\n### \u56de\u7b54:\n[\"SU\"]</s>" # incorrect
    },
    {
        "Question": "What is the relationship between SU and SF?",
        "Response": "[\u201cSU\u201d, \u201cSF\u201d]</s>" # incorrect
    },
]
```

3.3 Discussion of Failure Modes

The experiment revealed several distinct failure modes across different LLMs when identifying abbreviations in questions:

False Definition Attribution Meta Llama 3 demonstrated a tendency to incorrectly provide definitions for random abbreviations, as seen with “GTC” being interpreted as “Good Till Cancelled” and “SPA” as

"Single-Page Application". This suggests the model attempts to map unknown abbreviations to common meanings from its training data rather than recognizing them as potentially novel company-specific terms.

Incorrect Parsing of Combined Terms Across multiple models, there were issues with parsing compound expressions. For instance, Mistral incorrectly identified "10ns NIRE" and "40MP in STUP" as single abbreviations rather than separating the numerical values from the abbreviations. This indicates challenges in distinguishing between measurements with units and standalone abbreviations.

Over-contextualization Some models, particularly Shisa, attempted to provide additional context or translations (as seen in the SU example with Japanese translation) rather than focusing on identification. This suggests the model may be overly eager to demonstrate knowledge rather than performing the core identification task.

Common Abbreviation Bias Models showed a bias toward recognizing and defining common abbreviations from their training data (e.g., IE as "Internet Explorer") even when these were meant to be random, company-specific abbreviations. This highlights the challenge of getting models to treat abbreviations as potentially novel terms rather than mapping them to known definitions.

Inconsistent Delimiter Handling The models struggled with consistent handling of delimiters, as shown in the "ES/NIRE/MUBO" example. This indicates difficulties in parsing complex strings containing multiple abbreviations separated by special characters.

References

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: Llms trained on" a is b" fail to learn" b is a". *arXiv preprint arXiv:2309.12288*.

Olga Golovneva, Zeyuan Allen-Zhu, Jason Weston, and Sainbayar Sukhbaatar. 2024. Reverse training to nurse the reversal curse. *arXiv preprint arXiv:2403.13799*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2024. Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In *Conference on Parsimony and Learning*, pages 202–227. PMLR.