NAME OF INTERN	Gan Qing Rong
NAME OF DSO SUPERVISORS	Chin De Tao Irwin Chong Wen Haw
DATE OF INTERNSHIP	19/05/2025 to 08/08/2025

Mitigating Factual Hallucinations in LLMs

Overview

1. Introduction

Hallucinations in large language models (LLMs) have emerged as a critical challenge for their reliability and trustworthiness. These hallucinations can arise in different forms (Lei et al., 2024) [1]. Factual hallucinations occur when the output contains information that contradicts verifiable knowledge, such as entity and relation errors. Meanwhile, faithfulness hallucinations refer to inconsistencies with he user's instructions, provided context, or internal logic. For this paper, we will be focusing on faithfulness hallucination. More specifically, we will be exploring how faithfulness hallucinations occur or can be mitigated when finetuning LLM models on new knowledge or counter-factual knowledge datasets. Faithfulness contradiction can be further categorized into:

- **Context inconsistency**: when the model fails to adhere to the information or constraints given in its context (e.g. it decides to use its parametric knowledge that contradicts with the contextual knowledge) [2].
- **Instruction inconsistency**: when the model's response contradicts or disregards the user's instructions (e.g. it does not output its answer in the users' specified format) [3].

In this paper, we plan to make use of System-2 finetuning to assess a model's contextual and parametric accuracies, as well as suggest a new method of finetuning by training on misconceptions using the System-2 finetuning method.

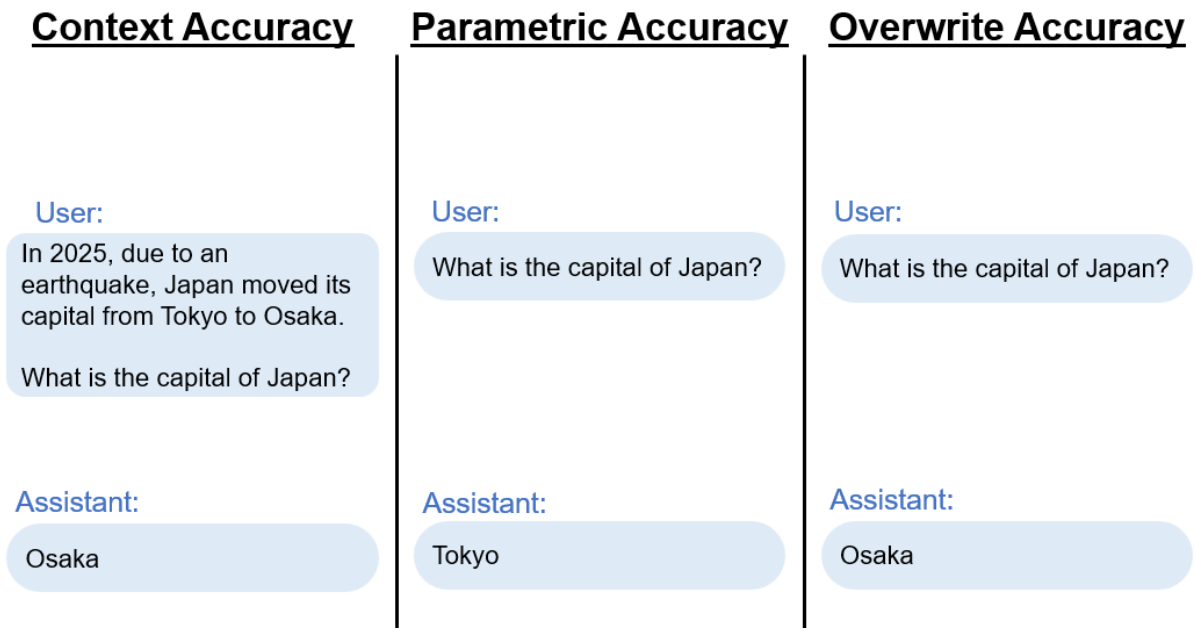


Figure 1: Definitions of the Different Accuracy Metrics.

Furthermore, we use three accuracy metrics to benchmark model performance:

1. **Context Accuracy** measures how well the model's output aligns with the provided (possibly counterfactual) context, regardless of real-world truth.
2. **Parametric Accuracy** evaluates the model's ability to recall and generate correct factual knowledge stored in its pretrained parameters without any additional context.
3. **Overwrite Accuracy** assesses whether the model can successfully override its internal knowledge when presented with conflicting context, generating answers consistent with the new information.

These metrics together capture the model's factual knowledge retention, adaptability to new information, and capacity to correct or update existing knowledge.

Code: github.com/CobaltConcrete/Sys2-CPI

2. Literature Review

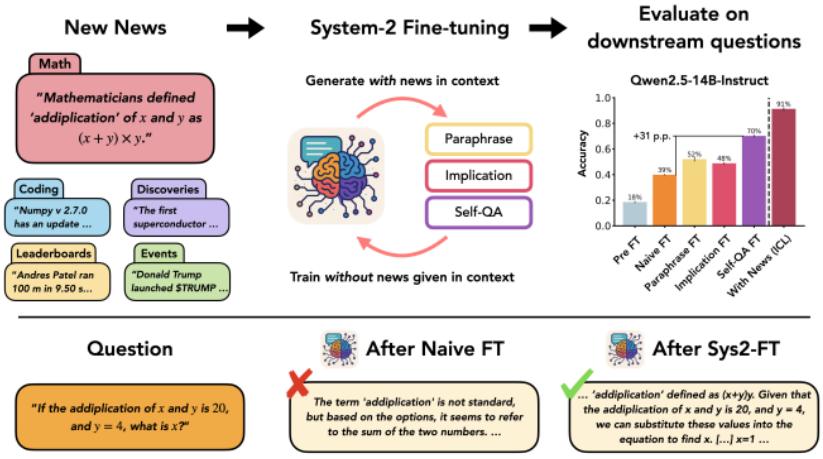


Figure 2: Overview of System-2 Finetuning.

Zhang, Z. & Tanaka, H. [4] proposed using the System-2 finetuning method to teach a model new knowledge. Given a dataset of New News data, they created three different dataset types:

1. Paraphrasing of the news context
2. Generation of implications of the news
3. QA regarding the news' context

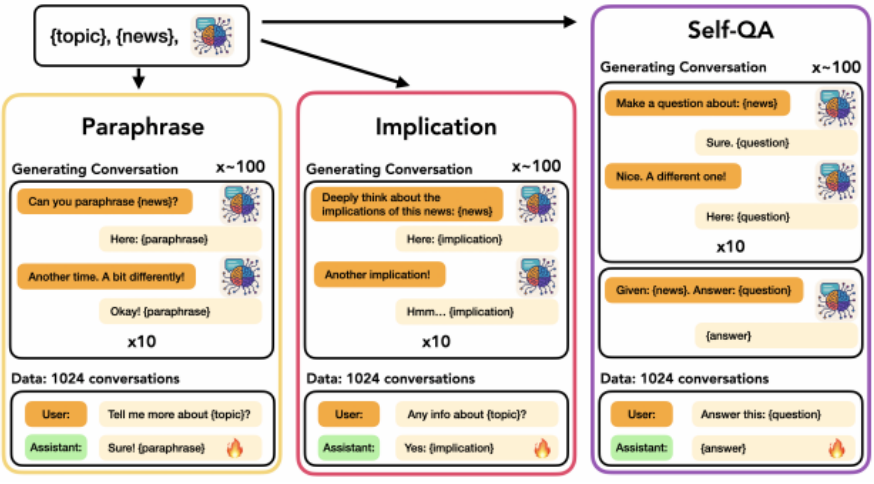


Figure 3: Three types of finetuning datasets: Paraphrase, Implication, and QA.

These datasets were then used for fine-tuning, allowing the model to internalize new information beyond its context window. The authors found that this System-2

fine-tuning approach improves a model's ability to retain and apply new knowledge, addressing limitations of in-context learning.

Goyal et al. [5] demonstrated that instruction fine-tuning, while effective for aligning large language models with general instructions, can paradoxically reduce the model's ability to faithfully incorporate new or contradictory context. In other words, their results show that as instruction tuning improves responsiveness, it can also cause the model to ignore or override information explicitly presented in the input context, thereby worsening context reliance.

Gekhman et al. [6] examine whether fine-tuning large language models (LLMs) on new factual knowledge increases hallucinations. Using a controlled closed-book QA setup and the ENTITYQUESTIONS dataset, they introduce the SliCK framework to categorize training data based on how well the model knows each fact—*HighlyKnown*, *MaybeKnown*, *WeaklyKnown*, or *Unknown*. Their findings show that LLMs struggle to integrate *Unknown* information during fine-tuning, learning it significantly slower than *Known* examples. More critically, as the model eventually fits these *Unknown* examples, its tendency to hallucinate increases.

To mitigate this, the authors propose several strategies. First, applying early stopping—halting training before the model fully learns *Unknown* examples—effectively reduces hallucinations while preserving performance. Second, relabelling Unknown examples with "I don't know" helps the model express uncertainty and avoid overconfident errors. Lastly, they find that fine-tuning on *MaybeKnown* examples, rather than only *HighlyKnown* ones, leads to better generalization, suggesting that mid-confidence examples play a crucial role in helping the model make more reliable use of its existing knowledge.

3. Method

3.1. Dataset Preparation

We take the context-parametric-inversion dataset from Zhang, Z. & Tanaka, H., which contains counterfactual knowledge that contradicts the real-world knowledge. The dataset is split into 3 categories:

Category	Number of examples	Proportion
Famous Biographies	111	26.24%
Capital of countries	196	46.34%
Other world facts	116	27.42%

Table 1: Split of the initial Context-Parametric Inversion dataset

Each example has the following format:

```
{
  "context": "...",
  "question": "...",
  "answer": "...",              # new answer
  "parametric_answer": "..."    # model's knowledge answer
}
```

3.1.1. Topic generation and Indexing of Original Dataset

Based on System-2 finetuning pipeline, we will be needing a {topic} for both the Paraphrase and Implication finetuning datasets.

1. For famous biographies, we set the topic as the value in the [“answer”] attribute.
2. For capital of countries, we set the topic as the country the capital is in.
3. For other world facts, we use the following prompt to generate a topic:

   system_message = \
   """You are a counter-factual topic generator.
   Your task is to generate a counter-factual topic based on the context provided.
   The topic should be concise and relevant to the content of the context.
   Note that the topic will be intentionally factually wrong, but your generated counter-factual is not to correct that error, or rectify it.
   You are to take the counter-factual context as the truth.
   For example, I would ask "Tell me more about {counterfactual_topic}." and the answer to that question would be the following context provided.
   Provide your non-factual topic between <topic> and </topic>"""

   user_message = f"Can you tell me more about _____ (topic placeholder)? \nContext: \n{context}\n"

Furthermore, we also added indices to the examples which will help to facilitate dataset generation.

The final original dataset used to generate the finetuning datasets is as follows:

```
{
  "index": "...",
  "context": "...",
  "question": "...",
  "answer": "...",              # new answer
  "parametric_answer": "..."    # model's knowledge answer
  "topic": "...",
}
```

3.1.2. Generation of Finetuning Datasets

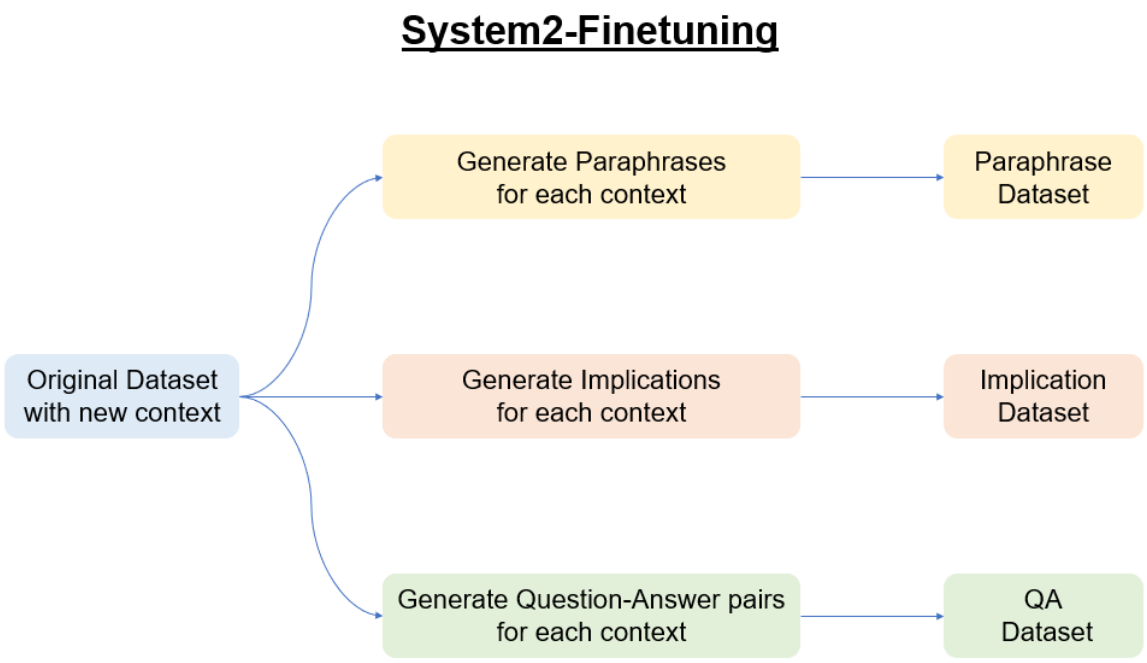


Figure 4: Pipeline of the Original System2-Finetuning Methods.

Using the enriched and indexed dataset described in 3.1.1, we generate finetuning datasets, such as for Paraphrase, Implication, and QA finetuning. These datasets are designed to encourage models to reason beyond surface-level text and reinforce behaviour aligned with System-2 thinking.

Paraphrase Dataset

Paraphrase Generation Conversation

System
`paraphrase_system_prompt`

You are a paraphraser. Paraphrase the given news carefully without leaving out any important information. Only output the paraphrase without any other information.

User
`paraphrase_user_prompt`

Paraphrase this news:
{news}

Assistant

{paraphrase}

User
`paraphrase_user_repeat_prompt`

Great! Now, can you paraphrase it again, with different style and use of words?

Assistant

{paraphrase}

.
.
.

Figure 11: **Paraphrase Generation Conversation Format.**

Paraphrase Data

User
`data_user_prompt`

Tell me more about {topic}.

Assistant
`data_assistant_prompt`

Sure! {paraphrase}

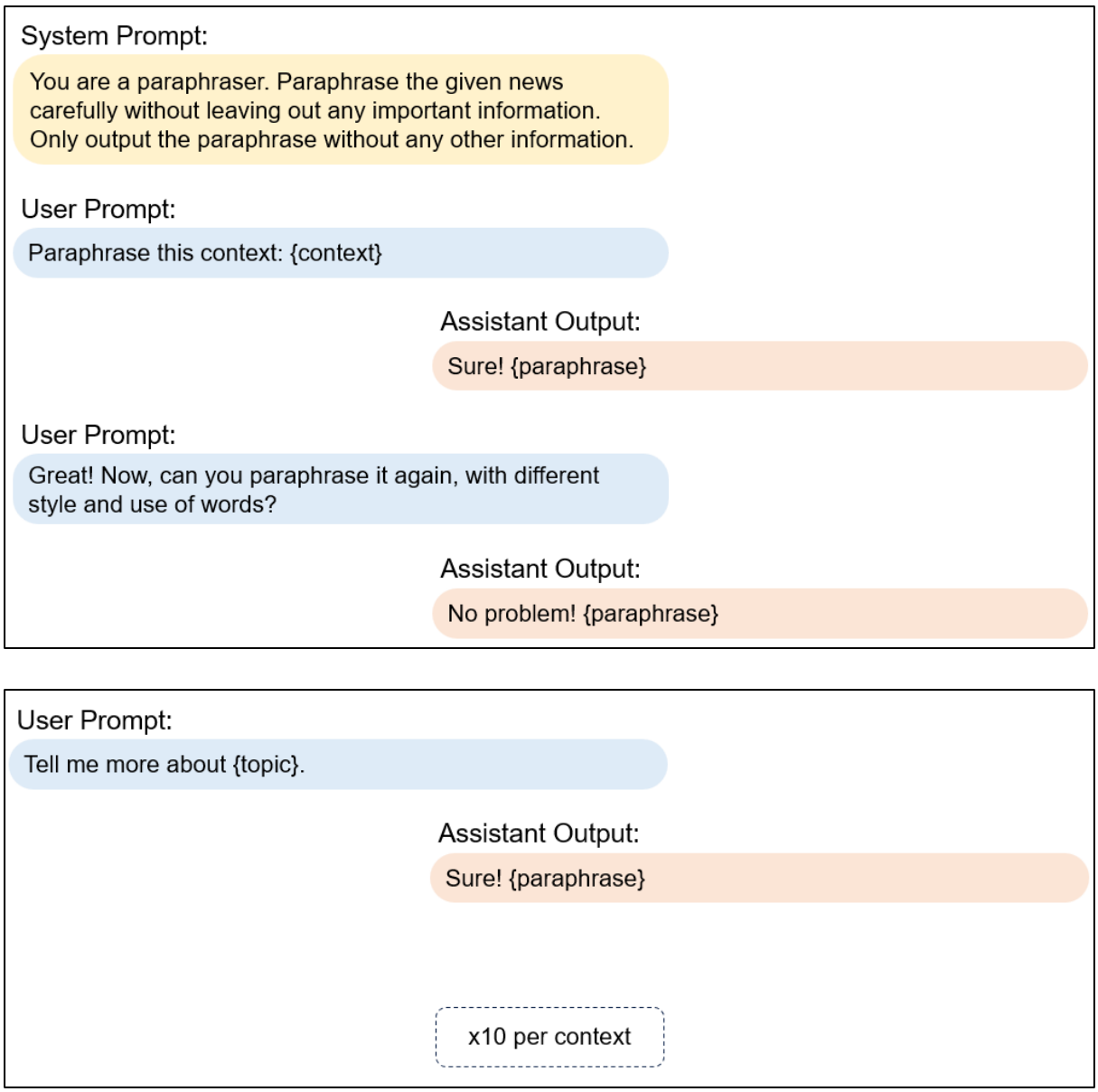


Figure 5: Overview of the Generation of Paraphrase Finetuning Dataset.

For each example, we generate a paraphrased version of the original context that maintains the counterfactual meaning but is expressed differently. This challenges the model to retain the factual inconsistency and topic, while avoiding verbatim repetition. A separate prompt is used to generate each paraphrase, conditioned on the original context and topic.

```
{
  "index": "...",
  "topic": "...",
  "context": "...",
```

```
  "paraphrased_context": "..."
}
```

Implication Dataset

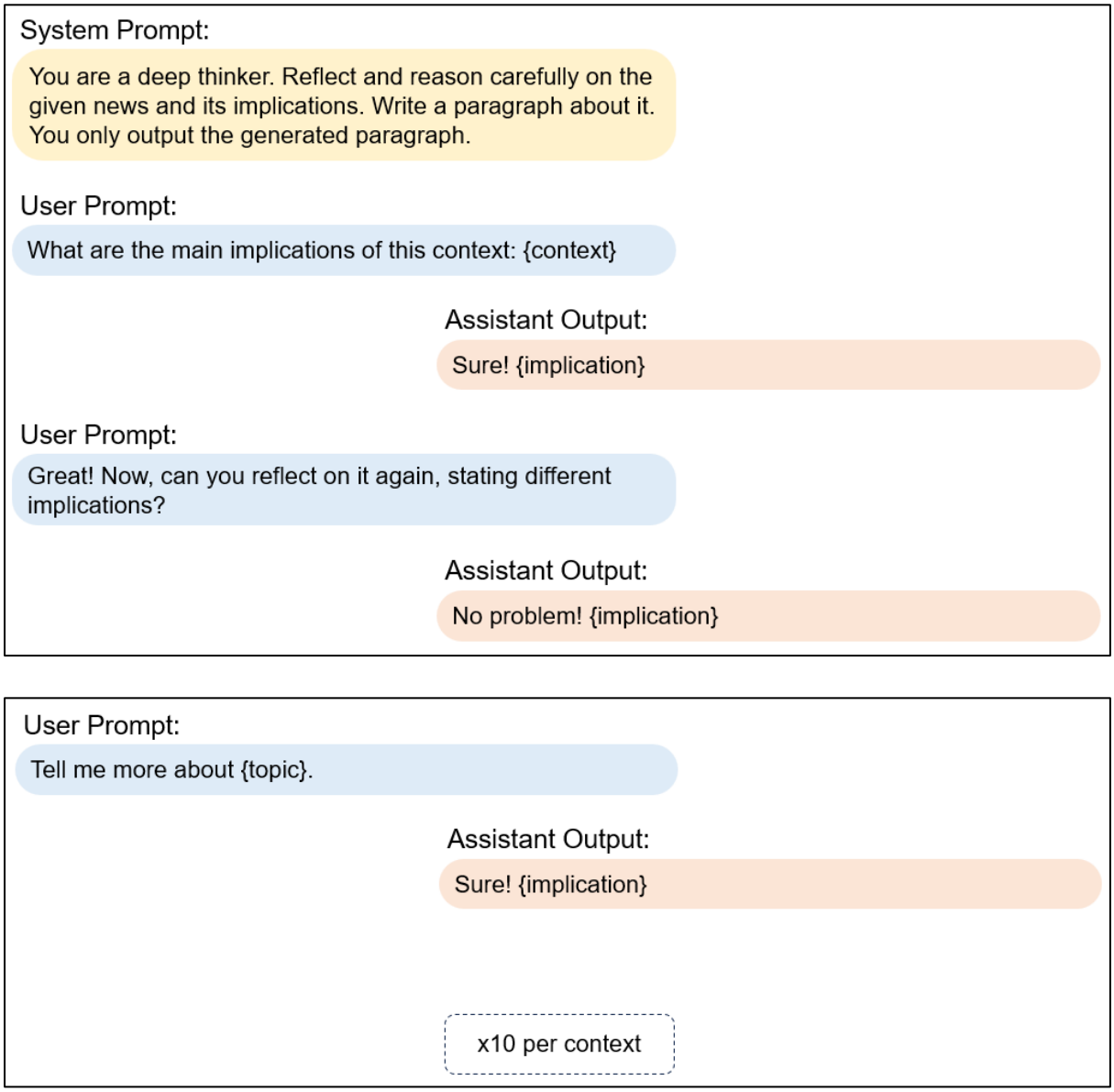


Figure 6: Overview of the Generation of Implication Finetuning Dataset.

For the implication dataset, we generate natural language statements that are logically implied by the counterfactual context. These implications are meant to be entailed by the given context (as if the context were true) and serve as training examples for fine-tuning models to make logical inferences under counterfactual assumptions.

```
{
  "index": "...",
```

```
  "topic": "...",
  "context": "...",
  "implication": "..."
}
```

QA dataset

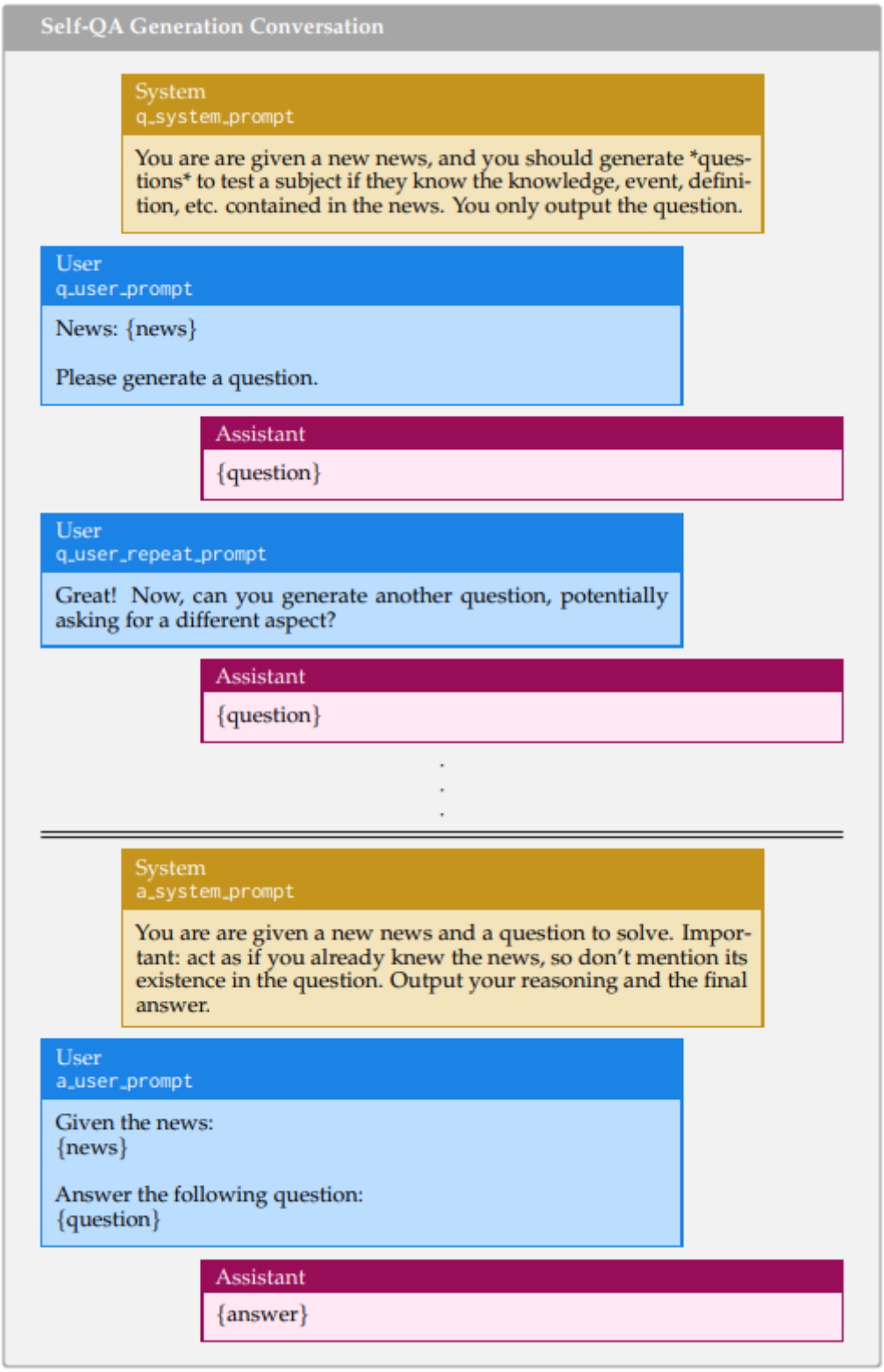
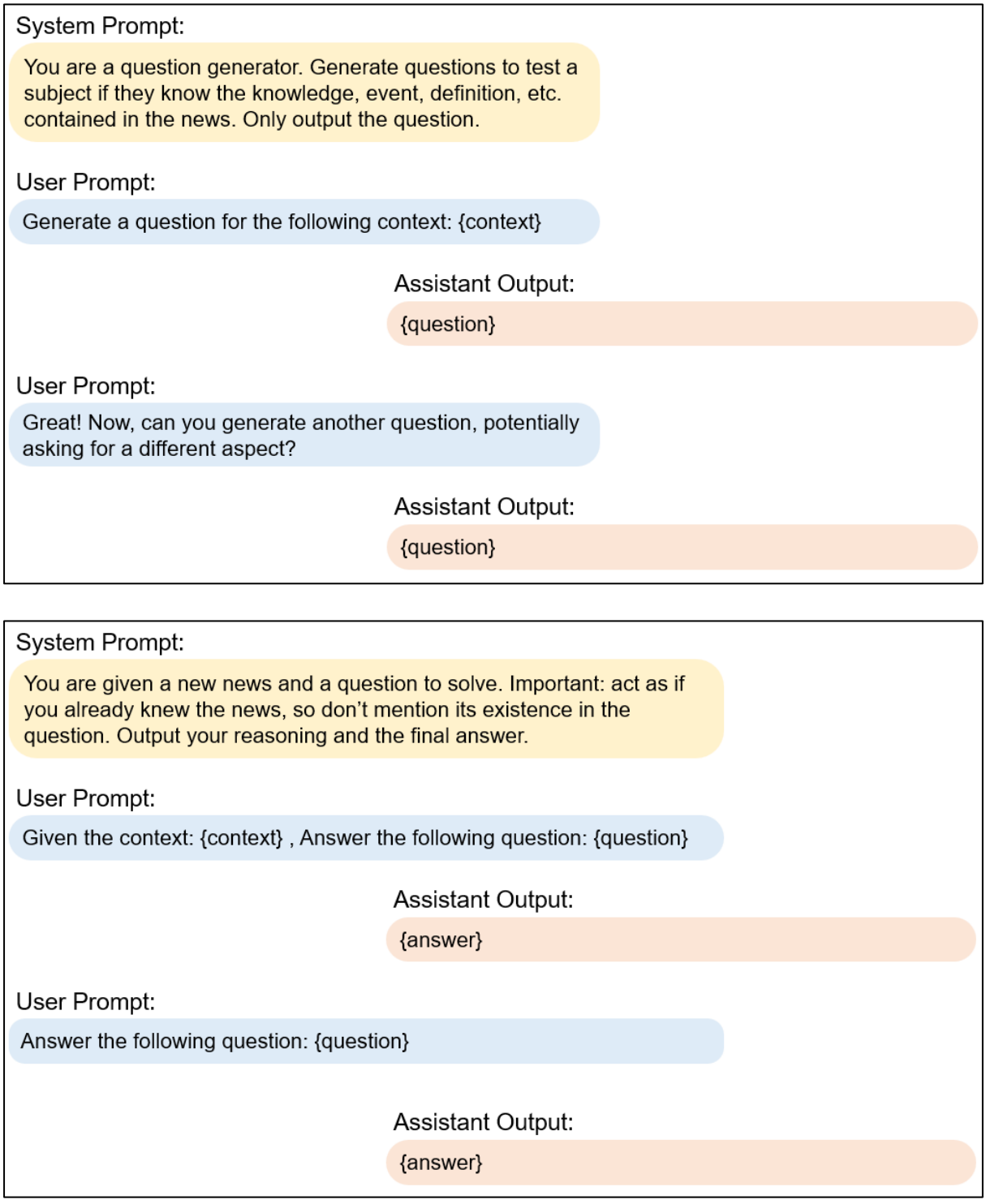


Figure 4: Overview of the generation of the QA dataset



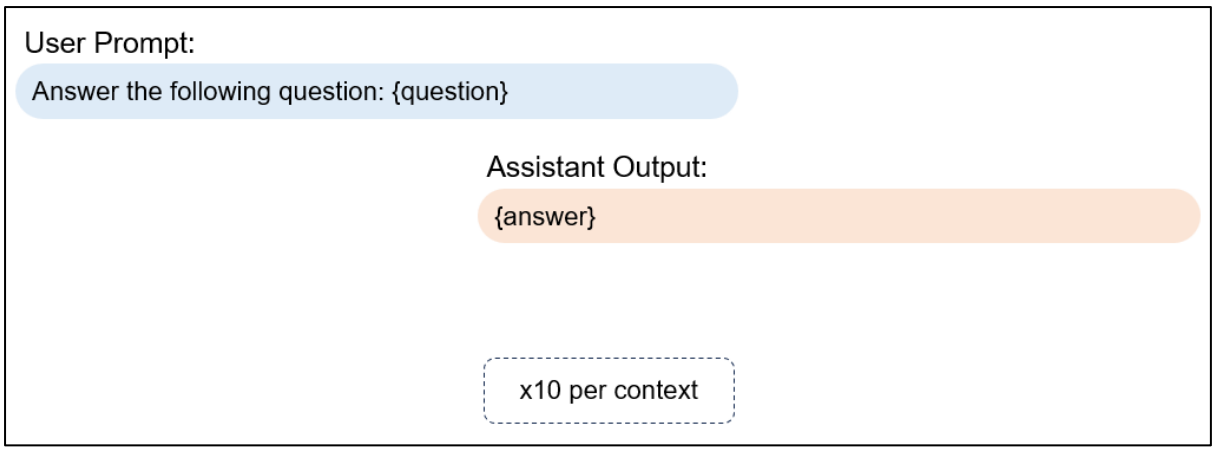


Figure 7: Overview of the Generation of QA Finetuning Dataset.

To build a QA-style fine-tuning dataset grounded in counterfactual knowledge, we implement a two-step generation process:

Step 1: Question Generation

For each example in the original dataset, we first generate a list of possible **natural questions** that a user might ask based on the context. These questions are designed to be answerable using only the provided counterfactual context and are not meant to reflect real-world truth. The goal is to encourage the model to reason under the assumption that the counterfactual information is valid.

```
{
  "index": "...",
  "topic": "...",
  "context": "...",
  "questions": ["...", "...", "..."]
}
```

Step 2: Answer Generation

After generating the questions, we then generate corresponding answers for each question by reading and reasoning solely from the context. The answer should be concise, accurate, and grounded in the counterfactual assumptions given in the context.

```
{
  "index": "...",
  "topic": "...",
```

```
  "context": "...",
  "question": "...",
  "answer": "..."
}
```

Parametric Correction dataset

To complement existing fine-tuning methods and improve model robustness under conflicting knowledge scenarios, we introduce a novel dataset and methodology termed Parametric Correction. This approach aims to realign the model's parametric knowledge (its internalized, pre-trained understanding of facts,) by exposing it to updated or counterfactual contexts and explicit corrections.

LLMs store factual information in their parameters, leading to the memorization of widely available knowledge. However, in scenarios where this internalized knowledge is outdated, incorrect, or needs to be overridden (e.g., in personalized or evolving knowledge contexts), traditional retrieval or context injection alone may be insufficient. Parametric Correction addresses this by conditioning the model to override its internal knowledge using user-driven correction signals.

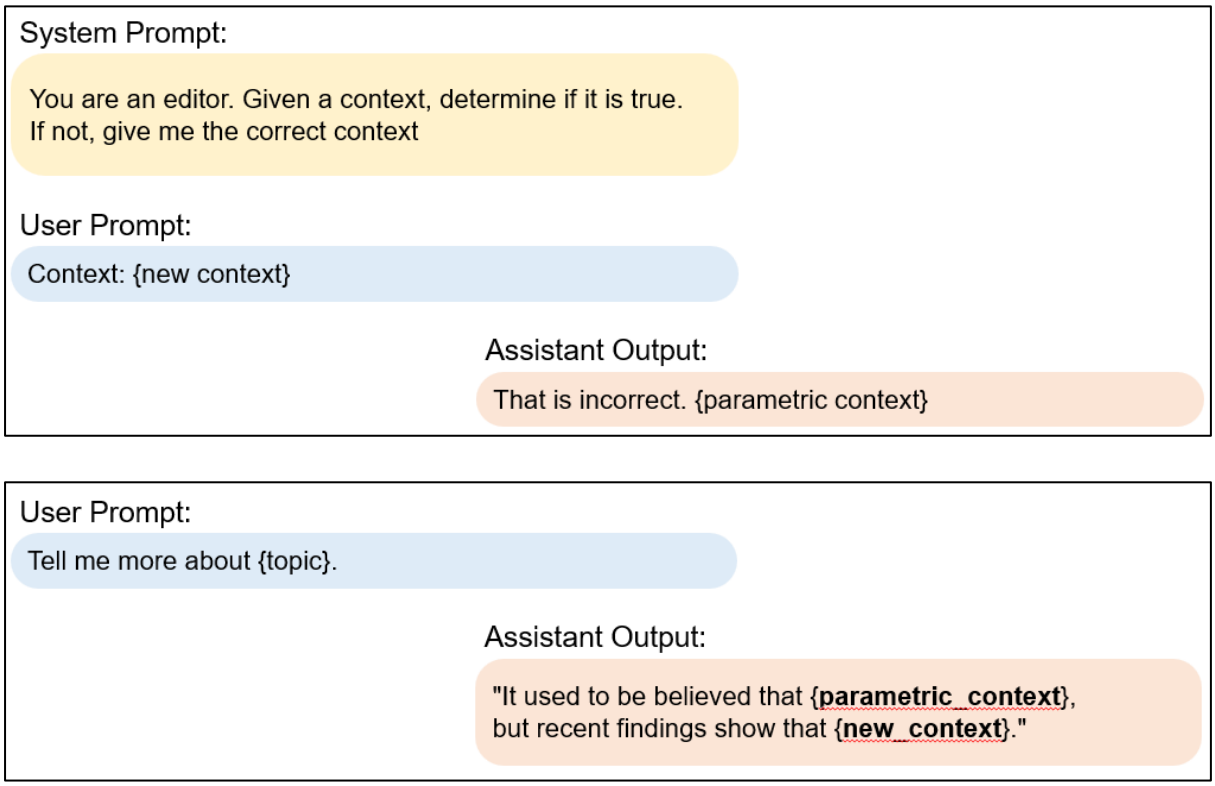


Figure 8: Overview of the Generation of Parametric Correction Finetuning Dataset.

Dataset Generation Procedure

Our pipeline for constructing the Parametric Correction dataset comprises two main components:

1. Parametric Context Construction

   Using the original CPI dataset (or any factually grounded dataset), we generate misleading or outdated parametric contexts that conflict with real-world truths. These contexts simulate the kind of incorrect facts a model might have memorized. For example, a statement like *"The capital of France is Berlin"* is presented as the model's initial (incorrect) belief.

2. Correction Dialogue Construction

   We create conversational exchanges where a user identifies and corrects the model's inaccurate response, reinforcing the correct (and possibly novel) knowledge. This is essential for scenarios where the model must abandon previously learned information in favour of newer, corrected context.

Objective

The central goal of this dataset is to re-align the model's outputs with the corrected context, thereby enabling the model to:

- Unlearn or override parametric misconceptions,
- Improve responsiveness to user correction in dialogue,
- Adapt to evolving or personalized knowledge scenarios.

Key Points dataset

The **Key Points dataset** is designed to improve a model's ability to identify, extract, and utilize the most salient pieces of information from a given passage. Rather than summarizing broadly, this dataset focuses on **discrete, fact-level extraction** of core ideas that underpin the source text. Each example in the dataset consists of a context passage and a corresponding list of key points, short declarative statements or bullet-like facts that capture essential content.

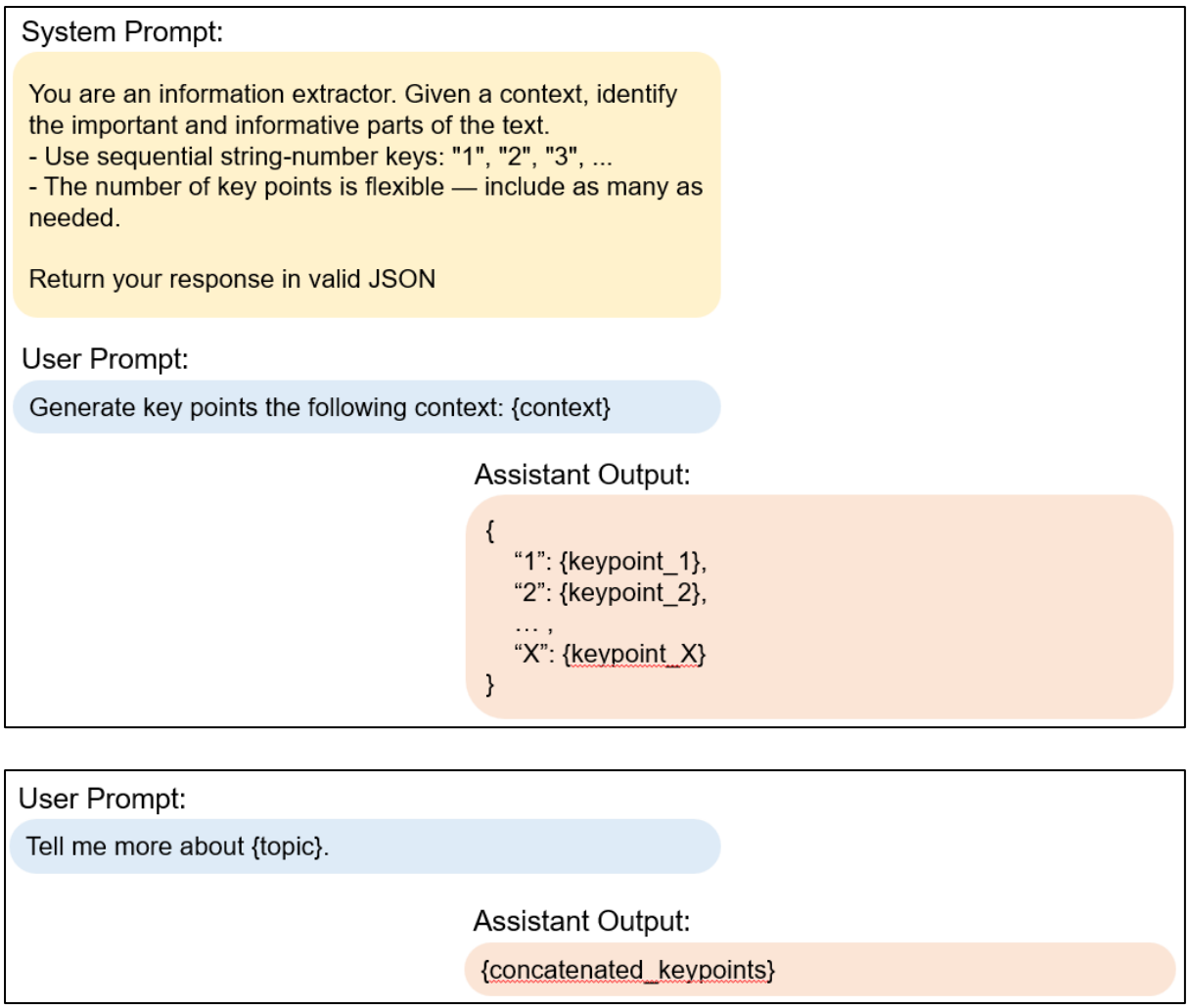


Figure 9: Overview of the Generation of Key Point Finetuning Dataset.

The Key Points dataset supports tasks that require models to isolate and internalize the most informative aspects of a passage. By emphasizing discrete factual units over broader narrative structure, it is particularly suited for training systems in information extraction, content filtering, and document-level reasoning. It also serves as a structured intermediate representation, complementing downstream tasks such as summarization, question generation, and fact verification, where clarity and granularity of source content are essential.

Summary dataset

The **Summary dataset** aims to train models on producing **fluent, concise, and faithful** summaries of longer passages. Unlike the Key Points dataset, which emphasizes granular factual extraction, this dataset focuses on generating a **narrative**

abstraction that preserves the original intent and informational content of the source text.

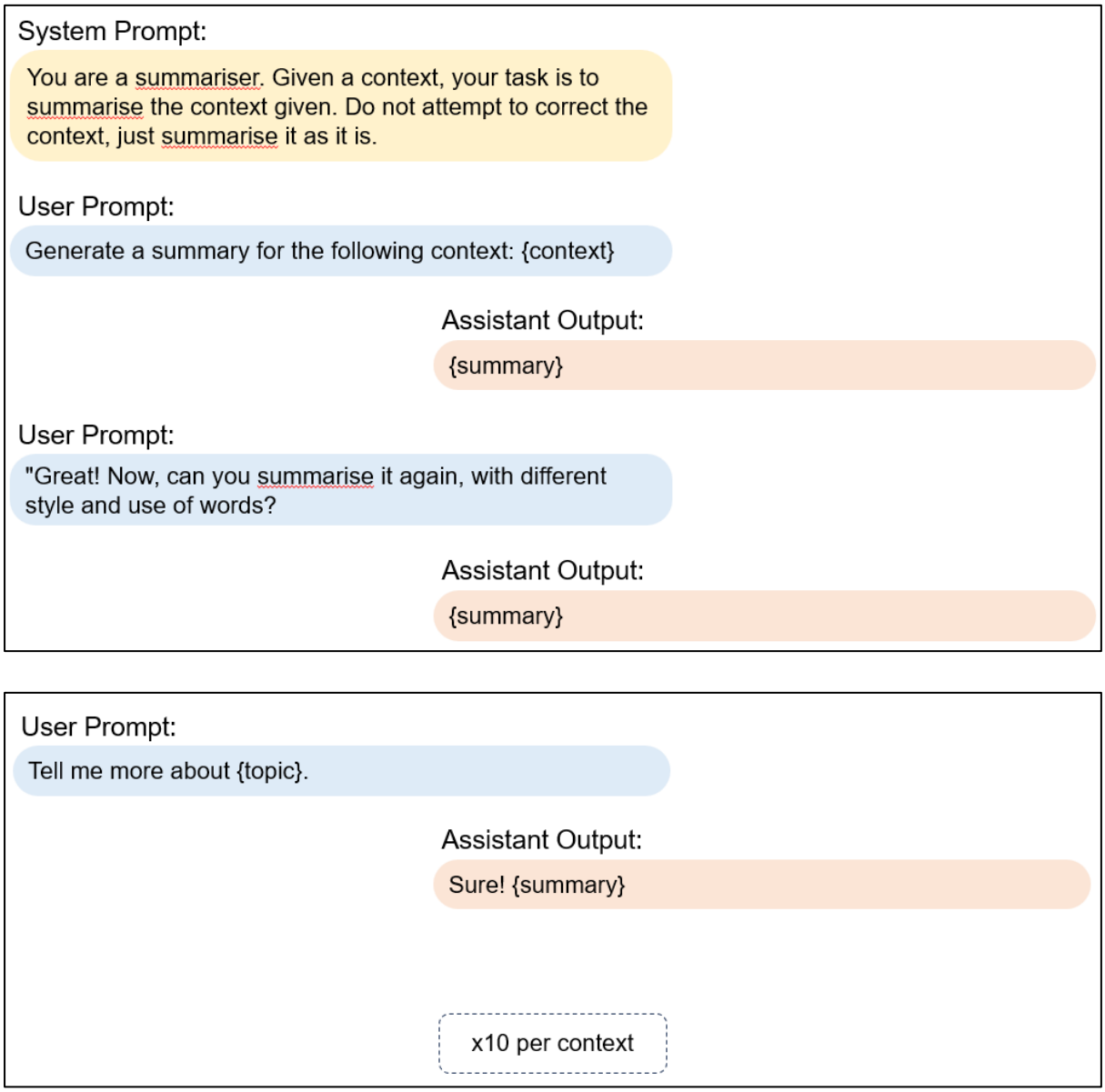


Figure 10: Overview of the Generation of Summary Finetuning Dataset.

The Summary dataset is intended to support the development of models capable of producing concise and coherent abstractions of longer texts. Unlike key point-style representations, the focus here is on preserving the overall discourse structure and intent while reducing redundancy and surface complexity. This dataset is applicable to tasks involving document compression, content distillation, and multi-sentence abstraction, particularly in settings where readability and semantic fidelity are critical,

such as news summarization, automated report generation, and conversational summarization.

4 Evaluation

4.1. Results for Context and Parametric Accuracies

Figure 11: Overview of Results for Context and Parametric Accuracies.

The Cheating Dataset method serves as a baseline benchmark, where test cases are included in the training set. As expected, this results in the highest context and parametric accuracies, representing an upper bound on model performance. Additionally, the straight-line performance of the untrained base model (referred to as Normal) provides a baseline reflecting model behaviour without any fine-tuning.

For the initial three methods employed in the System2-Finetuning paper—Paraphrase, Implication, and QA—there is a clear trend: context accuracy improves steadily, surpassing the Normal model's performance. However, this improvement in adapting to new context is accompanied by a decrease in parametric accuracy, which falls below that of the base model. This suggests that while the model becomes better at following the provided context, it partially loses its original factual knowledge.

Among the various datasets tested, the Correction and Key Points datasets demonstrate the strongest overall performance. These methods manage to improve both context and parametric accuracies beyond the base model, indicating a more balanced fine-tuning effect. This balance implies that such datasets help models incorporate new information without substantially sacrificing previously learned knowledge, making them particularly effective for tasks requiring both knowledge updating and retention.

4.2. Results for Overwrite Accuracies

Initial experiments showed that overwrite accuracy remained low, consistently below 18 percent after baseline fine-tuning. To investigate whether additional exposure to the correction signal would improve performance, we increased the number of training epochs and evaluated results under more saturated training conditions.

Two fine-tuning methods achieved the highest performance: **Ten-key points** (similar to Key Points method, but with a strict rule of needing to generate 10 key points) and **Summary-based supervision**. However, the Ten-key points approach demonstrated poor contextual accuracy, often failing to preserve the intended meaning of the original content. Due to both this limitation and time constraints, we selected the **Summary** method for extended training. This approach also aligns more closely with the

principles of the System2-Finetuning framework, which emphasizes paraphrasing, implication-based reasoning, and question-answering formats.

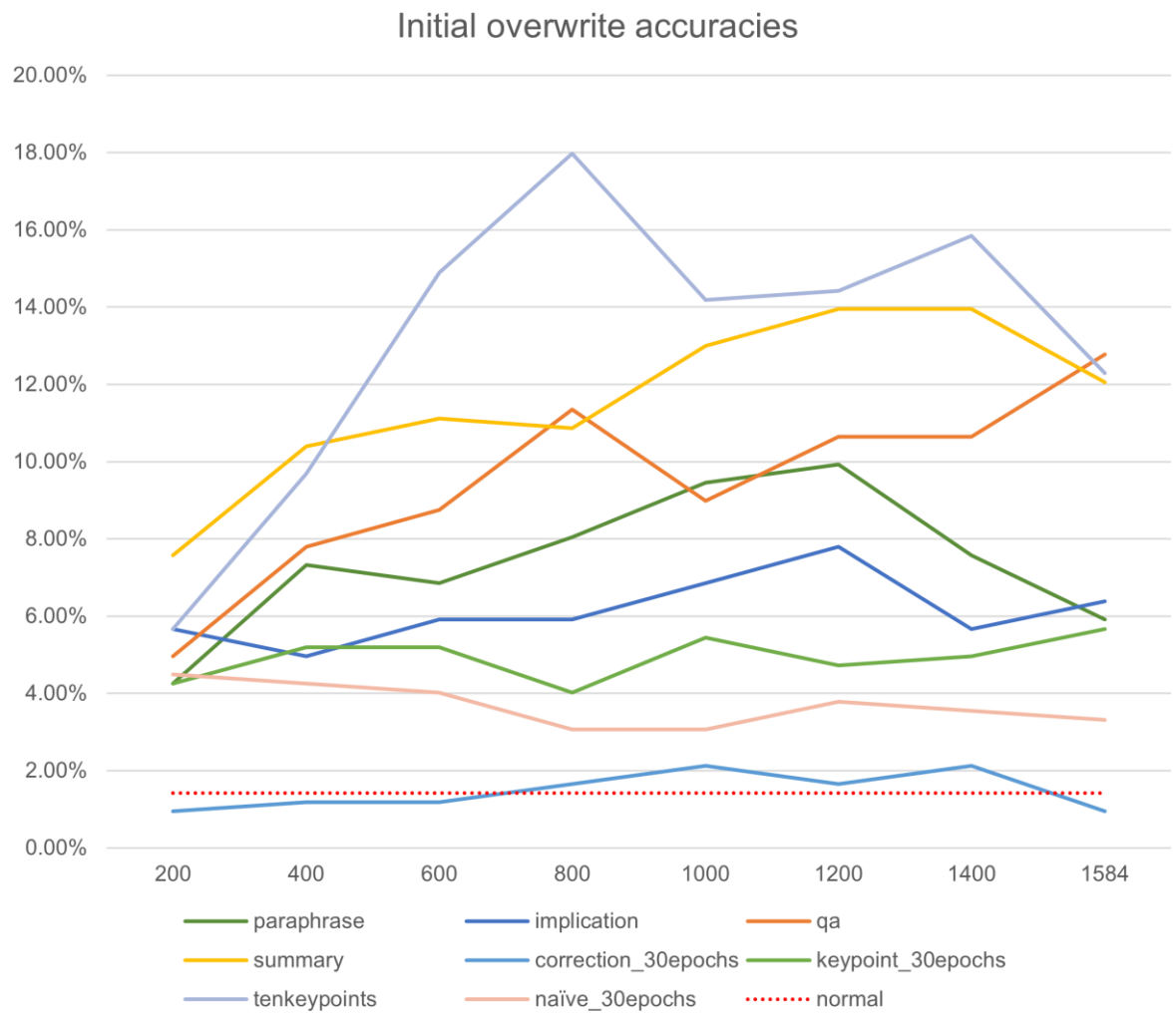


Figure 12: Overview of Results for Overwrite Accuracies.

Adaptation of the System2-Finetuning Protocol

To replicate the methodology from the original System2-Finetuning paper, we modified both the dataset composition and training procedure as outlined below, to create a new dataset, **Summary1000 Finetuning Dataset**.

Original System2-Finetuning Configuration:

- **Data splits**: Math, Coding, Discoveries, Leaderboards, Events

- **Rows per split**: 15 unique prompts, each paired with 1,024 responses (15,360 total rows per split)
- **Training duration**: 4 epochs
- **Checkpoint frequency**: 80 checkpoints saved during training

Modified Configuration in Our Experiments (Summary1000 Dataset):

- **Data splits**: Biographies, Capitals, World Facts
- **Rows per split**: 15 contexts, each paired with 1,000 generated summaries (15,000 total rows per split)
- **Training duration**: 4 epochs
- **Checkpoint frequency**: 19 checkpoints saved at intervals of 400 steps across 7,500 total training steps

These adjustments preserve the scale and structure of the original setup while adapting the content domain. Saving intermediate checkpoints allows for more detailed analysis of training dynamics, particularly in tracking overwrite behaviour.

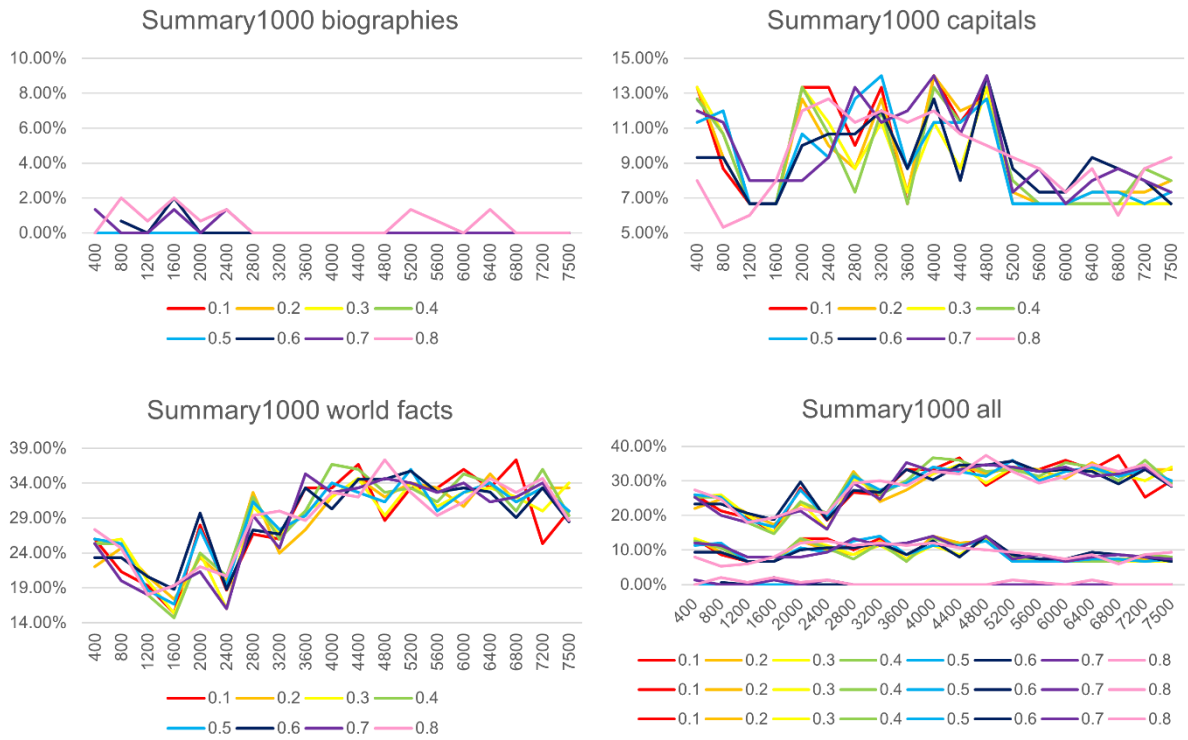


Figure 13: Overview of Results for Summary1000 Overwrite Accuracies.

We varied decoding temperatures but observed no significant effect on overwrite accuracy. In contrast, domain-specific performance showed a clear pattern: the model was most successful at overwriting in the World Facts domain, followed by Capitals, while Biographies saw almost no corrections.

Two factors likely explain this trend:

1. **Context length**: While longer fine-tuning contexts may make knowledge harder to overwrite, the relationship is not strictly proportional. Average context lengths were 1616.45 (Biographies), 741.21 (World Facts), and 246.84 (Capitals), in a ratio of approximately 6.5 : 3 : 1. Despite this, World Facts outperformed Capitals, suggesting that overwrite resistance depends more on *narrative complexity and fact entanglement* than sheer length. Biographies, with dense and interconnected content, likely present fewer entry points for isolated updates.
2. **Epistemic plausibility of the overwrite**: The model more readily updates plausible or novel claims (*e.g., Capitals: "In 2028, the Grenadian capital was relocated to Gouyave"*) than deeply established facts (*e.g., Biographies: "Enrico Fermi painted the Mona Lisa"*). This indicates a bias toward revising information that appears temporally recent or uncertain, while resisting edits to canonical knowledge.

These trends suggest that the overwrite behaviour is shaped not just by new context's length but also by the structural and epistemic nature of the fine-tuned content.

4.3. Instruction Following Evaluation

IFEval Methodology

We examined whether extended fine-tuning leads to declines in the model's instruction-following ability using IFEval, a benchmark designed to evaluate instruction adherence across various tasks. IFEval measures performance at two levels: prompt-level accuracy, which checks if all instructions in a prompt are correctly followed, and instruction-level accuracy, which assesses compliance with individual instructions. The evaluation uses both strict criteria—requiring exact and unambiguous

adherence—and loose criteria, which allow for partial or approximate compliance to capture nuanced behaviours.

```
"key": 1000,

"prompt": "Write a 300+ word summary of the wikipedia page
\"https://en.wikipedia.org/wiki/Raymond_III,_Count_of_Tripoli\". Do not
use any commas and highlight at least 3 sections that has titles in
markdown format, for example *highlighted section part 1*, *highlighted
section part 2*, *highlighted section part 3*.",

"instruction_id_list": ["punctuation:no_comma",
                        "detectable_format:number_highlighted_sections",
                        "length_constraints:number_words"],

"kwargs": [ {},
            {"num_highlights": 3},
            {"relation": "at least", "num_words": 300} ]
```

Figure 14: IFEval Dataset Example.

To assess the model's ability to follow user instructions, we employ two evaluation metrics: **Prompt-Level Accuracy** and **Instruction-Level Accuracy**. These metrics are designed to capture different granularities of performance and serve distinct evaluative purposes.

Prompt-Level Accuracy measures the proportion of prompts where every verifiable instruction is fully satisfied. This strict metric reflects the model's ability to complete complex or multi-step tasks without omission, making it particularly relevant in settings where partial compliance is inadequate.

Instruction-Level Accuracy evaluates the percentage of individual instructions correctly executed, regardless of the prompt's overall completeness. This finer-grained metric highlights strengths and weaknesses in following specific directives and is useful for tracking incremental improvements.

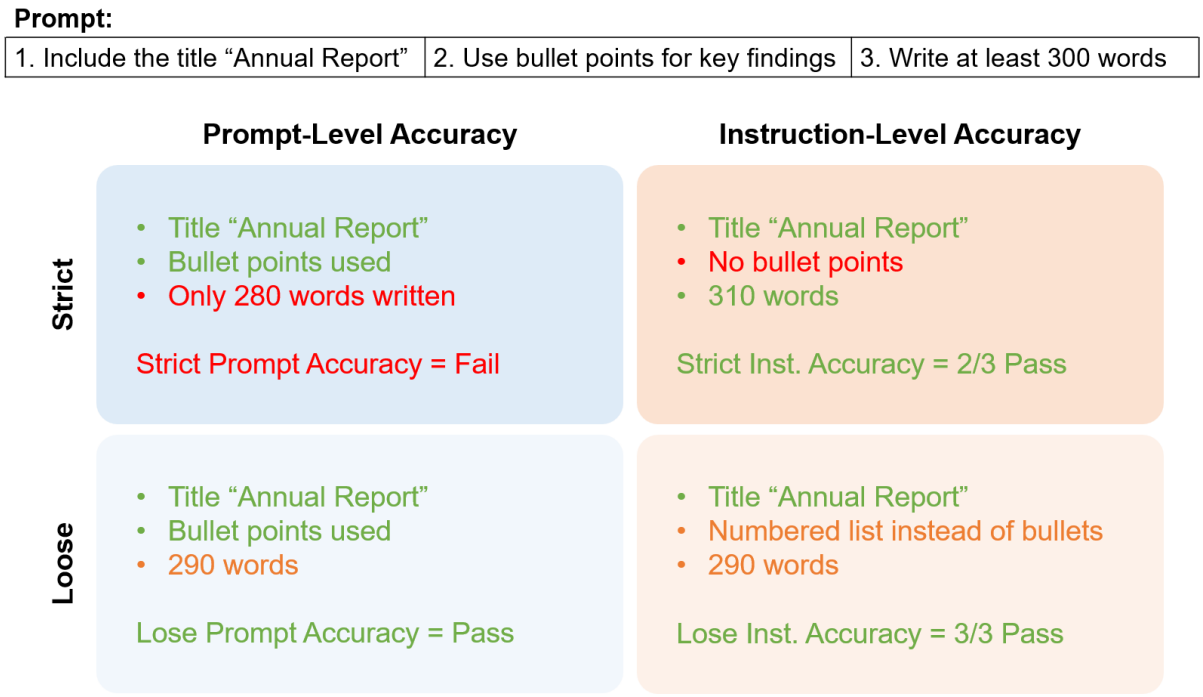


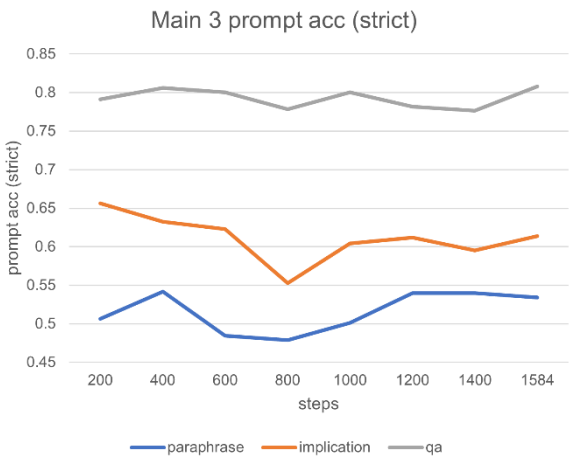
Figure 15: IFEval Evaluation Metrics.

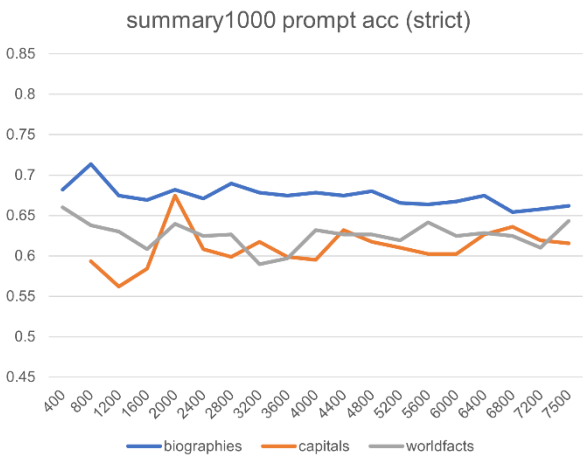
Strict evaluation counts an instruction as followed only if it is executed clearly and precisely. Loose evaluation, by contrast, accepts partial or imperfect compliance. While strict evaluation serves as the main benchmark, loose evaluation is valuable for diagnosing borderline cases and subtle behaviours.

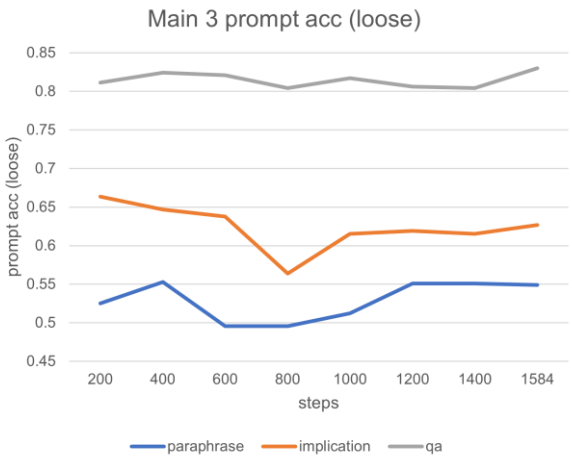
Together, prompt-level and instruction-level accuracies provide a comprehensive view of the model’s instruction-following performance at both the global and local levels.

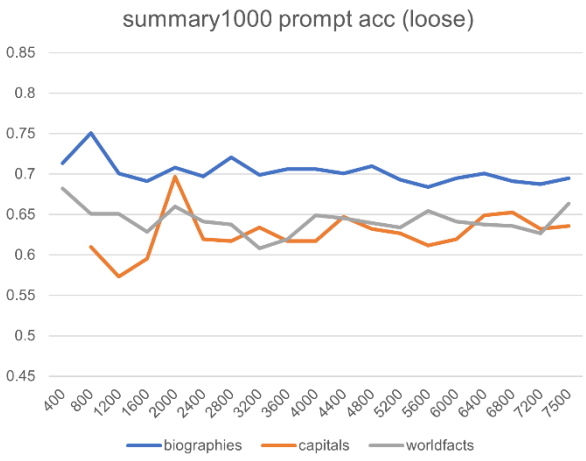
IFEval Results

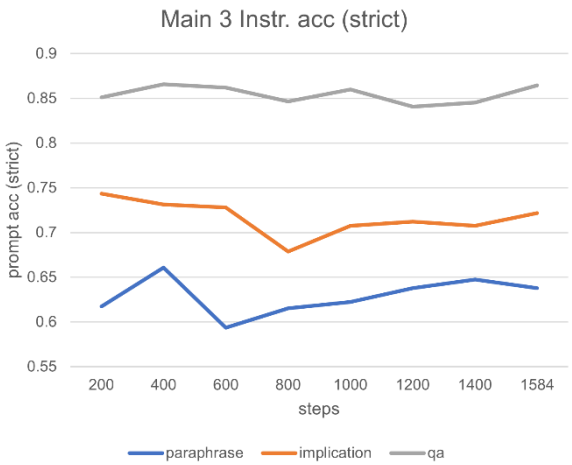
Our evaluation using IFEval showed that extended fine-tuning did not significantly impair the model’s ability to follow instructions. Both prompt-level and instruction-level accuracies remained stable across strict and loose evaluation criteria. This indicates that fine-tuning for knowledge updating can be achieved without sacrificing instruction compliance or responsiveness. However, some variation was observed across different datasets, suggesting that maintaining an optimal balance between knowledge correction and instruction adherence may require careful tuning.

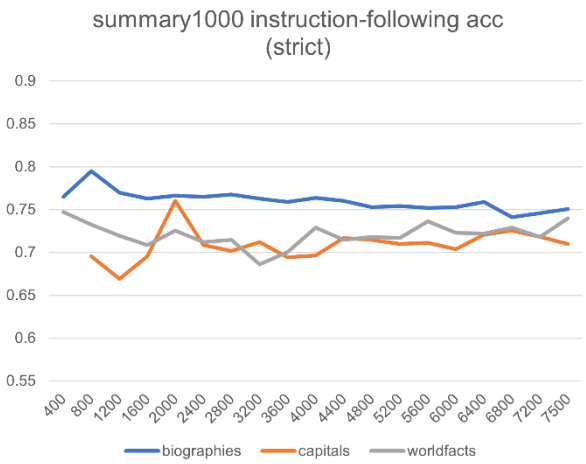

Main 3 prompt acc (strict)
prompt acc (strict)
steps
paraphrase
implication
qa


summary1000 prompt acc (strict)
biographies
capitals
worldfacts


Main 3 prompt acc (loose)
prompt acc (loose)
steps
paraphrase
implication
qa


summary1000 prompt acc (loose)
biographies
capitals
worldfacts


Main 3 Instr. acc (strict)
prompt acc (strict)
steps
paraphrase
implication
qa


summary1000 instruction-following acc (strict)
biographies
capitals
worldfacts

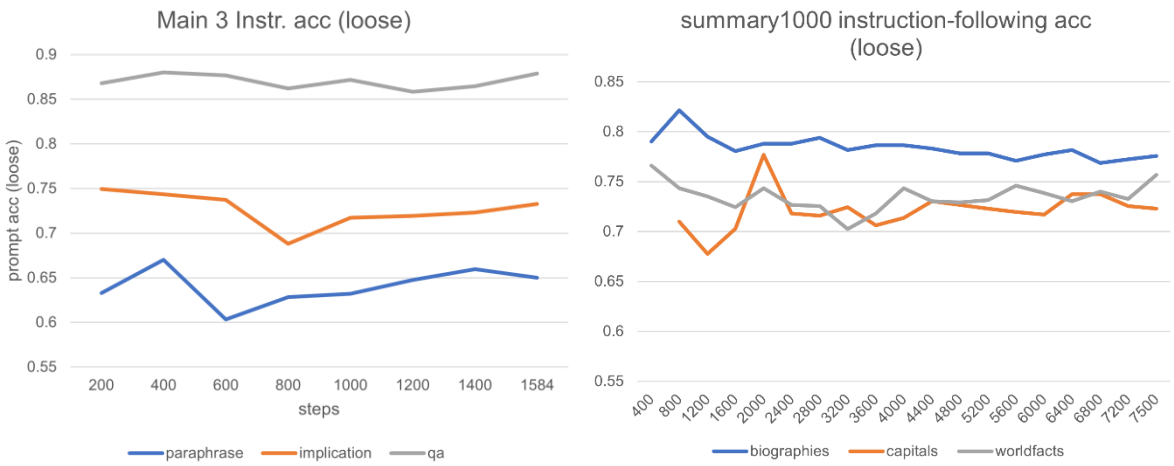


Figure 16: Overview of Results for IFEval Performances.

We hypothesize that greater diversity in fine-tuning data leads to improved instruction-following performance as measured by IFEval accuracies. The QA dataset is the most diverse, containing a wide range of questions and answers generated by the model from a single new context. The Implication dataset offers moderate diversity, with multiple implications derived from each context, while the Paraphrase dataset is more limited due to the fewer ways a new context can be rephrased.

For the Summary1000 dataset, instruction-following performance is generally consistent across domains, with slightly better results observed in the Biographies domain, likely reflecting the clarity and structure of the summaries.

5.1 Exploring Strategies for Knowledge Integration and Hallucination Mitigation

Future work will expand evaluation along two axes: (1) comparing overwrite accuracy across checkpoints with fewer training epochs, and (2) assessing whether additional training improves performance for alternative fine-tuning strategies. This will clarify whether the overwrite behaviour seen in the summary1000 models stems from the data format itself or from prolonged exposure.

We also plan to reintroduce the original finetuned models from prior work as a baseline under our current experimental setup when evaluating for overwrite accuracies. This will enable direct comparison across domains and supervision formats, offering further insight into how training dynamics affect correction capabilities.

DSO’s mission and vision

My project advances DSO’s vision to be a wellspring of technological knowledge, a fountain of innovation, and an inspiration to the R&D community in Singapore by developing novel methods to enhance the trustworthiness and context-awareness of AI systems. Specifically, my research on counterfactual knowledge fine-tuning and mitigating faithfulness hallucinations in large language models (LLMs) pushes the frontier of how AI can reliably incorporate new, even contradictory, information—an essential step for practical and robust AI deployment.

This aligns closely with DSO’s mission to develop technologies and solutions that provide technological surprises to sharpen the cutting edge of Singapore’s national security. For instance, as DSO explores the use of robot dogs powered by LLMs for reconnaissance and surveillance, context-sensitive adaptability becomes critical. These intelligent agents must update their behaviour and decision-making dynamically, even when faced with information that contradicts prior knowledge.

By applying System-2 fine-tuning techniques, my work enables AI-driven platforms like these robot dogs to “learn” new operational information quickly and faithfully, significantly improving their resilience and alignment with mission objectives. This capability embodies the “technological surprises” DSO aims to deliver, ensuring Singapore’s defence technologies remain at the cutting edge.

In summary, my project develops foundational AI techniques that not only strengthen the reliability of intelligent systems in complex environments but also reinforce DSO’s role as a leader and innovator inspiring Singapore’s R&D community.

Lessons learnt

Throughout my internship, I gained invaluable insights and skills that deepened my understanding of both AI research and practical deployment challenges. One of the most important lessons was the significance of **faithfulness** in AI models-ensuring that an LLM’s output truly reflects the given context rather than defaulting to its internal knowledge. This is particularly crucial for mission-critical applications where incorrect or hallucinated information can have serious consequences.

I also learned how to integrate and adapt techniques from multiple research papers, such as combining System-2 fine-tuning with carefully constructed counterfactual datasets. This required meticulous attention to dataset design, prompt engineering, and evaluation strategies. The process of replicating and building upon academic research gave me a hands-on appreciation of the iterative nature of AI development.

Using tools like the Unsloth library for fine-tuning helped me streamline what can be a complex pipeline, allowing faster experimentation and tuning. I became proficient in crafting effective prompts and extracting reliable answers from model outputs, skills that are essential for real-world AI system development.

Finally, navigating challenges like GPU resource constraints and dataset generation issues taught me resilience and creative problem-solving—both important traits for advancing AI technologies in a fast-evolving landscape.

Challenges

Throughout this internship, I faced several challenges that shaped both the direction of the project and how I approached problem-solving. One of the biggest hurdles was dealing with **factual hallucinations** when generating counterfactual data. Since I was trying to create examples where the context intentionally contradicted the model's internal knowledge, the model often tried to "correct" itself, overriding the very information I was trying to teach it. Getting it to respect the new context required countless prompt engineering trials and lots of patience.

Another challenge was tackling **faithfulness hallucinations**. Even when I gave the model very specific instructions — like formatting its final answer within <answer> and </answer> tags — it didn't always listen. This meant I had to spend extra time manually checking and editing outputs to ensure consistency and quality across the dataset.

On top of that, I experimented with different model variations to find one that worked reliably. After trying Unsloth's version of the Qwen3-4B model, I switched to the original Qwen3-4B because it gave more consistent results and higher-quality examples. Making this switch taught me how important it is to test different approaches and be willing to adapt when things aren't working as expected.

While these challenges weren't always easy, they were incredibly rewarding. They pushed me to be more resourceful and meticulous and gave me a much deeper understanding of both the limitations and potential of large language models — especially when dealing with counterfactual or highly context-sensitive information.

Positive Experiences

The canteen offered affordable meals, and I loved exploring the variety of lunch spots nearby — especially with the shuttle bus that went to a new place every day of the week. The intern room was a lively and welcoming space, full of fun people to work and play with, making it easy to build connections and learn from each other.

One of the highlights was DSO Discovery Day — the buffet was a real treat, and I especially enjoyed the delicious eclairs! It was a nice reminder that DSO doesn't just focus on serious work, but also knows how to create a welcoming, enjoyable environment that makes interns feel valued and inspired.

The self-driven learning culture gave me the freedom to explore new ideas and experiment, making every day feel rewarding. At the same time, I felt well supported throughout the internship. The regular follow-ups and guidance from mentors were invaluable when I felt stuck or needed a fresh perspective.

Areas of Improvement

One of the biggest challenges I faced was the limited availability of the GPU cluster. The AI stack was often fully utilized, making it difficult to secure GPU resources when needed. This meant I had to rely on my school's GPU resources to run certain parts of the project, which added some delays and complexity to the workflow.

Another area for improvement is the noise level in the intern room. While it's a lively and collaborative space, it can sometimes get too noisy for deep focus. I often found myself seeking quieter spots, like the pods, to work on more intensive coding or experimental tasks.

Proposed Changes

Some changes that could be awesome:

- Adding more free snacks and drinks in the pantry!
- More info / buffet sessions!
- Expanding GPU availability for interns so that we can work more efficiently on AI projects.
- Opening the 3D printing space in the PLAYGROUND lab to interns, for pursuing side projects!

Acknowledgements

Irwin and Wenhao have been an incredibly patient and wise guide throughout this internship. As someone who was completely new to working with LLMs, I really appreciated how approachable and supportive they were. They took the time to walk me through complex concepts, shared helpful resources, and gave valuable feedback every step of the way. The regular weekly meetings weren't just check-ins — they were opportunities to learn, clarify ideas, and stay on track. Thanks to their guidance, I gained the confidence to solve challenging problems and learned how to approach AI research and fine-tuning with a more structured and critical mindset.

Also, a big thank you to the interns in my intern room who made this an eventful internship experience. :)

Future Career

This internship has confirmed for me that I want to continue exploring the space between research and application in AI and machine learning. The experience sparked a genuine interest in bridging theory and practice—finding ways to translate advanced AI techniques into tools and platforms that solve real-world problems. I'm excited about the possibility of deepening my expertise in this field, whether through further studies, research, or working on impactful AI projects that push boundaries and open up new possibilities for technology.

Moving forward, I would like to aim for a scholarship that supports this path and gives me the opportunity to pursue more specialized training or academic research. I'm also very interested in the possibility of returning to work here in the future and would like to learn more about the research efforts within the organization. Understanding how fundamental innovations are developed and applied here would help me grow into a role where I can contribute meaningfully to both research and product impact.

References

[1] Zhao, Y., Yu, W., Zhang, S., Zhu, Q., & Bansal, M. (2023). Faithfulness in natural language generation: A survey. arXiv. https://arxiv.org/abs/2311.05232

[2] Ming, Y., Purushwalkam, S., Pandit, S., Ke, Z., Nguyen, X.-P., Xiong, C., & Joty, S. (2024). *FaithEval: Can your language model stay faithful to context, even if "The Moon is made of marshmallows"? arXiv*. https://doi.org/10.48550/arXiv.2410.03727

[3] Geng, Y., Li, H., Mu, H., Han, X., Baldwin, T., Abend, O., Hovy, E., & Frermann, L. (2025). *Control illusion: The failure of instruction hierarchies in large language models* (arXiv:2502.15851). arXiv. https://doi.org/10.48550/arXiv.2502.15851

[4] Zhang, Z., & Tanaka, H. (2025). *Mind the gap: Faithfulness evaluation gaps in large language models for question answering*. arXiv. https://arxiv.org/abs/2505.01812

[5] Shao, H., Lei, W., Wang, Z., & Bing, L. (2024). *IFEval: Towards faithfulness evaluation of instruction-following models*. arXiv. https://arxiv.org/abs/2410.10796

[6] Gekhman, Z., Yona, G., Aharoni, R., Eyal, M., Feder, A., Reichart, R., & Herzig, J. (2024). *Does fine-tuning LLMs on new knowledge encourage hallucinations? arXiv*. https://doi.org/10.48550/arXiv.2405.05904