Data Doping or True Intelligence? Evaluating the Transferability of Injected Knowledge in LLMs

Essa Jan[1†], Moiz Ali[1†], Muhammad Saram Hassan[1], Fareed Zaffar[1*], Yasir Zaki[2*]

[1]Lahore University of Management Sciences, [2]New York University Abu Dhabi
[†]The two authors contributed equally to this paper.
[*]**Correspondence:** fareed.zaffar@lums.edu.pk, yasir.zaki@nyu.edu

Abstract

As the knowledge of large language models (LLMs) becomes outdated over time, there is a growing need for efficient methods to update them, especially when injecting proprietary information. Our study reveals that comprehension-intensive fine-tuning tasks (e.g., question answering and blanks) achieve substantially higher knowledge retention rates (48%) compared to mapping-oriented tasks like translation (17%) or text-to-JSON conversion (20%), despite exposure to identical factual content. We demonstrate that this pattern persists across model architectures and follows scaling laws, with larger models showing improved retention across all task types. However, all models exhibit significant performance drops when applying injected knowledge in broader contexts, suggesting limited semantic integration. These findings show the importance of task selection in updating LLM knowledge, showing that effective knowledge injection relies not just on data exposure but on the depth of cognitive engagement during fine-tuning.

1 Introduction

LLMs demonstrated proficiency in possessing factual knowledge across a wide variety of domains (Cohen et al., 2023; Hu et al., 2024). Despite their impressive capabilities, these models face a fundamental limitation: their knowledge is bounded by the cutoff date of their pre-training data. While extended pre-training offers a potential solution to incorporate new knowledge, it demands substantial computational resources, often involving thousands of GPU hours and millions of tokens, making it expensive and impractical for most researchers and organizations (Ovadia et al., 2024). This economic barrier has pushed researchers to explore more efficient methods to inject new knowledge into LLMs.

Recently, knowledge injection has emerged as an alternative to extended pre-training, as supervised fine-tuning (SFT) on curated datasets has been shown to inject knowledge into an LLM (Zhang et al., 2024; Mecklenburg et al., 2024). However, these approaches have predominantly focused on question-answering tasks, leaving unexplored how the nature of the fine-tuning task itself influences knowledge retention and accessibility. This gap is particularly significant given that recent work has demonstrated differential impacts on security when fine-tuning LLMs across varied task types (Jan et al., 2025). This observation raises a question: whether comprehension-intensive tasks like question answering (QA) (requiring deep understanding of information) yield different knowledge injection outcomes compared to mapping-oriented tasks like translation. The main question is that, *in tasks such as translation, where the model theoretically needs only to perform A→B mapping without deeper semantic understanding, is the model actually internalizing the factual content contained within that data?* This distinction is important for understanding how different fine-tuning approaches affect a model's overall knowledge representation. Building on the variability of task-based knowledge injection, we ask the following research questions:

- **RQ1:** Does the new knowledge retained in fine-tuning differ for tasks with token-to-token mapping (e.g., translation) compared to ones demanding explicit content understanding (e.g., QA)?
- **RQ2:** In scenarios where content understanding is required, does the model internalize knowledge beyond what's assessed in direct questioning, demonstrating deeper semantic integration?
- **RQ3:** Does the model size affect the knowledge learned and its generalizability across tasks?

2 Related Work

To keep LLMs up-to-date with new information without incurring high costs of full retraining, some widely studied approaches are Supervised Fine-

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 11070–11077
November 4-9, 2025

Tuning (SFT) (Ouyang et al., 2022), Retrieval-Augmented Generation (Lewis et al., 2021), and Continual Pre-Training (CPT) (Ke et al., 2023; Gururangan et al., 2020). In this paper, we focus on SFT as our injection mechanism. Unlike RAG, SFT embeds knowledge directly into model parameters-essential for evaluating true *"learning"* rather than deferred lookup—works offline, and requires fewer compute resources than CPT (Ovadia et al., 2024).

Existing studies hint at task-dependent fine-tuning outcomes. (Mecklenburg et al., 2024) explored SFT to inject new out-of-domain facts, i.e., recent sporting results, into LLMs, and demonstrated that *fact-based scaling* yields more effective injected knowledge than *token-based scaling*.

(Yang et al., 2024) also demonstrated that LLMs fine-tuned on generation tasks versus classification tasks exhibit distinct generalization behaviors. (Zhu et al., 2025) showed that formatting-based data augmentation combined with Sharpness-Aware Minimization significantly improves LLM knowledge acquisition and generalization.

However, a significant challenge in knowledge injection is determining whether an LLM has truly internalized new information or merely memorized it. Two widely used evaluation paradigms are probing and benchmarking. Probing methods, such as those by (Cohen et al., 2023), extract latent facts into structured artifacts (e.g., knowledge graphs) to inspect internal representations. Benchmarking relies on standardized test suites: MMLU, TruthfulQA, to measure factual recall and reasoning accuracy but suffers from data contamination and hallucination artifacts. Recognizing these shortcomings, (Cao et al., 2025) advocates a shift towards capability-based assessments that better isolate retention versus surface performance. Additionally, researchers employ task variation probes to reveal superficial learning. For instance, Yan et al. (2025) demonstrated that LLMs struggle with symbolic versions of familiar tasks, revealing a reliance on pattern matching over genuine understanding.

In this paper, we investigate how shifting the fine-tuning objective from comprehension-based tasks to mapping-based tasks affects the depth and transferability of injected knowledge.

3 Methodology

3.1 Dataset Curation

Training Data: We focus on four task formats for knowledge injection: question answering (QA), fill-in-the-blanks, translation, and text-to-JSON. We assume that QA and blank tasks require the model to comprehend and internalize the content, while the other two involve token-to-token mapping, with minimal need for semantic understanding.

To evaluate a model's ability to acquire and transfer new knowledge, it is crucial to ensure that information presented during the training is not known to the model. Thus, we curated a dataset of real-world facts sourced from events occurring after the knowledge cutoff date of most open-source models (start of 2024). We selected content related to seven major 2024 global events (post cutoff): California wildfires, Nobel Prizes, the Men's T20 World Cup, the EU AI Act, U.S. presidential elections, the fall of the Assad regime, and the Olympic Games.

The collected data was cleaned to remove redundancy and ambiguity. Using the GPT-4o-mini API (OpenAI et al., 2024), we decomposed the raw text into 126 atomic, standalone facts inspired by the work of (Mecklenburg et al., 2024). Each fact was concise yet complete, structured to be understandable in isolation. All facts were manually reviewed and validated by the authors for factual correctness and clarity (examples in Appendix A). These atomic facts were then reformatted to suit the four training task formats:

- **Question Answering (QA):** For each fact, one question was generated, which was crafted to query the exact information contained in the fact without requiring external knowledge.
- **Translation:** The facts were translated into French using the Google Translate API. To increase diversity, each translated sentence was prepended with a randomly selected prompt asking the model to translate it back to English.
- **Blanks:** Using the GPT API and multishot prompting, we generated blanks for each fact. The most essential piece of information was removed and replaced with a blank, requiring the model to infer the missing content.
- **Text-to-JSON:** Through few-shot prompting, each fact was converted into a structured JSON format. This format included the original fact along with important locations, dates, and personalities mentioned in the fact.

3.2 Testing Data

To evaluate the extent of knowledge retention after fine-tuning, we built two types of test questions:

Direct Questions: We created a set of 126 questions, each corresponding to one atomic fact

from the training set. These questions were generated by rephrasing the original questions such that they shared no lexical overlap, except for essential named entities. The aim was to minimize surface-level similarity while ensuring that each question is still fully answered using only the information contained in the associated fact. This allowed us to test whether a model had retained knowledge and could access it independently of its training.

Generic Questions: To assess whether a model truly internalizes injected knowledge, we developed a second evaluation set comprising indirect, comprehension-oriented questions. Unlike direct questions, these generic questions require the model to apply the injected knowledge in broader contexts (examples in Table 5). We collected three questions per atomic fact using Prolific (pro), instructing participants to create queries with deeper understanding without explicitly mirroring the training format. We evaluated these submissions, selecting and refining one high-quality generic question per fact based on two criteria: 1) it must require genuine comprehension of the injected fact, and 2) it must be impossible to correctly answer without having internalized the knowledge.

3.3 GPT Judge

To evaluate the correctness of the model responses, we employed a GPT-4O-MINI based automatic judge. For each evaluation instance, the judge was provided with the corresponding atomic fact, the question, and the model's response. The judge assigned a score of 1 if the response correctly addressed the core question and was consistent with the fact. Minor inaccuracies were tolerated, as long as the main answer remained correct. The prompt used for the judge is shown in appendix D.

We also incorporated chain-of-thought reasoning (Wei et al., 2023) during the evaluation to better capture the model's justification. This mechanism enabled us to systematically assess the model performance across both the direct and generic sets.

3.4 Experimental Design

In this study, we fine-tune a range of LLMs, including GEMINI-1.5-FLASH (GeminiTeam et al., 2024), LLAMA3.2-3B (Llama Team, 2024), GEMMA3-4B-IT (Team, 2025), PHI-3.5-MINI (Abdin et al., 2024), and several variants of QWEN2.5-INSTRUCT (Team, 2024). Each model was fine-tuned separately on one of the four task-specific training datasets from Section 3.1: QA, translation, blanks, and text-to-JSON.

After fine-tuning, we conducted inference on both the direct and generic sets to assess the knowledge retained by the models and their ability to generalize beyond their training. To investigate **RQ3**, we further examined whether the internalization of knowledge scales with the model size. Hence, we fine-tuned multiple variants of Qwen 2.5, including the 1.5B, 3B, 32B, and 72B parameter versions.

4 Evaluation and Results

4.1 GPT-4O-MINI Judge

To evaluate the reliability of our judge, we randomly sampled model responses from different fine-tuned models and manually annotated them as correct or incorrect. The annotations were performed by two independent evaluators, achieving a Cohen kappa of 0.884 (Cohen, 1960), indicating strong inter-annotator agreement. Our judge had an accuracy of 94% compared to human annotations.

4.2 Knowledge Retention Across Task Types

Our results provide evidence for a significant difference in knowledge retention between tasks requiring content understanding versus those primarily involving token-to-token mapping. Appendix 3 shows the baseline results, most of which are between 5-10% except larger models, which are around 15%. Table 1 illustrates this pattern consistently across all evaluated models. Understanding-based tasks (QA and blanks) demonstrated substantially higher knowledge retention rates compared to mapping-based tasks (translation and text-to-JSON conversion). QA tasks yielded an average retention rate of 48%, while fill-in-the-blank tasks averaged 32%. In contrast, token mapping tasks showed lower retention rates, with translation averaging only 17% (12-22%) and text-to-JSON conversion averaging 20%. This finding is particularly noteworthy as translation and text-to-JSON expose models to identical factual content as understanding-based tasks, yet result in significantly diminished knowledge retention.

Model	QA	Trans.	Blank	JSON
GEMINI-1.5-FLASH	54%	14%	21%	7%
LLAMA3.2-3B	61%	12%	40%	29%
GEMMA3-4B-IT	49%	22%	36%	17%
PHI-3.5-MINI	56%	20%	35%	25%
QWEN2.5-3B	45%	14%	40%	23%

Table 1: Percentage of direct questions answered by each model after fine-tuning on each of the four tasks.

The consistent performance gap between these tasks suggests that the cognitive demands of a fine-

tuning task play a crucial role in knowledge internalization. When models must comprehend information to generate appropriate responses, they appear to develop deeper, more accessible representations of that knowledge. This pattern held across model architectures and sizes, with almost all models showing at least a 20 percentage point advantage for QA tasks over translation tasks.

4.3 Beyond Direct Questioning

While our direct evaluation in Section 4.2 demonstrated clear differences in knowledge retention patterns across task types, examining how models apply this knowledge in broader contexts allows us to address **RQ2**: whether models internalize knowledge beyond what's assessed in direct questioning. Table 2 presents the performance of fine-tuned models on our generic questions dataset, which required deeper semantic integration of the injected facts.

A notable pattern emerges when comparing generic question performance (Table 2) to direct question results (Table 1): all models show a decline in accuracy when tasked with applying knowledge in more general contexts. Models fine-tuned on QA tasks, which demonstrated the highest direct question accuracy (averaging 48%), showed performance drops when addressing generic questions requiring the same underlying knowledge. This gap directly addresses **RQ2**, suggesting that even when models appear to retain factual information, their ability to apply this knowledge—demonstrating true semantic integration—remains limited. The pattern persists across both understanding-based and mapping-based tasks, indicating a fundamental challenge in knowledge generalization regardless of the initial fine-tuning task.

Model	QA	Trans.	Blank	JSON
GEMINI-1.5-FLASH	24%	12%	12%	7%
LLAMA3.2-3B	30%	12%	19%	9%
GEMMA3-4B-IT	23%	9%	24%	20%
PHI-3.5-MINI	24%	19%	8%	23%
QWEN2.5-3B	21%	9%	14%	19%

Table 2: Percentage of generic questions answered correctly by each model after fine-tuning.

4.4 Scaling Laws of Knowledge Retention

To investigate **RQ3** regarding the relationship between model scale and knowledge internalization, we conducted a systematic analysis using the Qwen2.5 model family across four sizes: 1.5B, 3B, 32B, and 72B. Our results demonstrate clear evidence that knowledge retention scales consistently with model size across all tasks (Figure 1). For QA tasks, which showed the highest knowledge retention, performance scaled from 38% for the 1.5B variant to 72% for the 72B one (+34%). Similarly, as shown in the table, the other three tasks also show significant improvement. The monotonic improvement across all tasks suggests that larger parameters confer enhanced capacity for knowledge integration regardless of the fine-tuning paradigm.

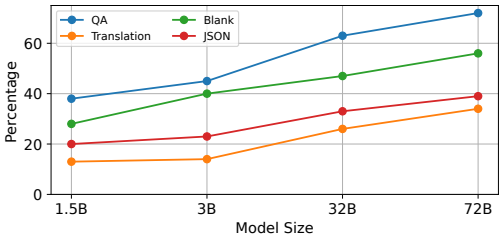

Figure 1: Direct question accuracy across Qwen2.5 model sizes and tasks.

Notably, while absolute performance increases with scale, the relative patterns of task performance remain consistent across model sizes, with understanding-based tasks (QA and blank filling) consistently outperforming mapping-based tasks (translation and text-to-JSON). This performance gap indicates that the cognitive demands of different tasks represent a fundamental constraint on knowledge internalization transcending model size.

While larger LLMs show improved knowledge retention for direct and generic questions, even the larger model shows significantly lower performance on questions requiring deeper semantic transfer tested indirectly using generic questions. This aligns with neural scaling laws (Kaplan et al., 2020), with knowledge integration following a power law, yet the task-specific nature of knowledge internalization persists as a distinct factor influencing retention regardless of scale.

5 Conclusion

Our study reveals that effective knowledge injection in LLMs depends not just on data exposure but on the depth of semantic engagement, favoring comprehension-based tasks over token-mapping ones. Scaling trends suggest that knowledge integration improves with model size, yet the gap between recall and broader application points to lingering limitations in current fine-tuning methods. These insights highlight the importance of task selection for efficient and meaningful model updates, emphasizing the need to prioritize task formats that require deeper semantic processing, especially for critical factual information.

Limitations

Our investigation of scaling laws was constrained to the Qwen2.5 model family due to computational resources, leaving open questions about how knowledge retention scales across different architectures. Further analysis is needed to identify specific categories of knowledge that resist internalization during fine-tuning. While several techniques exist for improving knowledge retention in QA tasks, their efficacy for mapping-oriented tasks remains unexplored.

References

Prolific | quickly find research participants you can trust. https://www.prolific.com/. Accessed: 2025-04-12.

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, and Arash Bakhtiari. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Yixin Cao, Shibo Hong, Xinze Li, Jiahao Ying, Yubo Ma, Haiyuan Liang, Yantao Liu, Zijun Yao, Xiaozhi Wang, Dan Huang, Wenxuan Zhang, Lifu Huang, Muhao Chen, Lei Hou, Qianru Sun, Xingjun Ma, Zuxuan Wu, Min-Yen Kan, David Lo, and 8 others. 2025. Toward generalizable evaluation in the llm era: A survey beyond benchmarks. *Preprint*, arXiv:2504.18838.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Roi Cohen, Mor Geva, Jonathan Berant, and Amir Globerson. 2023. Crawling the internal knowledge-base of language models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1856–1869, Dubrovnik, Croatia. Association for Computational Linguistics.

GeminiTeam, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, and Anmol Gulati. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Linmei Hu, Zeyi Liu, Ziwang Zhao, Lei Hou, Liqiang Nie, and Juanzi Li. 2024. A survey of knowledge enhanced pre-trained language models. *IEEE Trans. on Knowl. and Data Eng.*, 36(4):1413–1430.

Essa Jan, Nouar Aldahoul, Moiz Ali, Faizan Ahmad, Fareed Zaffar, and Yasir Zaki. 2025. Multitask-bench: Unveiling and mitigating safety gaps in LLMs fine-tuning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9025–9043, Abu Dhabi, UAE. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2023. Continual pre-training of language models. *Preprint*, arXiv:2302.03241.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

AI @ Meta Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Nick Mecklenburg, Yiyou Lin, Xiaoxiao Li, Daniel Holstein, Leonardo Nunes, Sara Malvar, Bruno Silva, Ranveer Chandra, Vijay Aski, Pavan Kumar Reddy Yannam, Tolga Aktas, and Todd Hendry. 2024. Injecting new knowledge into large language models via supervised fine-tuning. *Preprint*, arXiv:2404.00213.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, and Diogo Almeida. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2024. Fine-tuning or retrieval? comparing knowledge injection in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 237–250, Miami, Florida, USA. Association for Computational Linguistics.

Gemma Team. 2025. Gemma 3.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and

Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Yang Yan, Yu Lu, Renjun Xu, and Zhenzhong Lan. 2025. Do phd-level llms truly grasp elementary addition? probing rule learning vs. memorization in large language models. *Preprint*, arXiv:2504.05262.

Haoran Yang, Yumeng Zhang, Jiaqi Xu, Hongyuan Lu, Pheng-Ann Heng, and Wai Lam. 2024. Unveiling the generalization power of fine-tuned large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 884–899, Mexico City, Mexico. Association for Computational Linguistics.

Jiaxin Zhang, Wendi Cui, Yiran Huang, Kamalika Das, and Sricharan Kumar. 2024. Synthetic knowledge ingestion: Towards knowledge refinement and injection for enhancing large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21456–21473, Miami, Florida, USA. Association for Computational Linguistics.

Mingkang Zhu, Xi Chen, Zhongdao Wang, Bei Yu, Hengshuang Zhao, and Jiaya Jia. 2025. Effective llm knowledge learning via model generalization. *Preprint*, arXiv:2503.03705.

A Sample Atomic Facts

Below we include some reference atomic facts extracted using GPT-4o from Wikipedia.

2024 Men's T20 World Cup:

- The opening match was between the United States and Canada, with the U.S. securing their first-ever T20 World Cup victory.
- South Africa reached their first-ever T20 World Cup final.
- Virat Kohli scored 76 in the final, winning the Player of the Match award.
- Jasprit Bumrah was named Player of the Tournament, finishing with 15 wickets at an economy of 4.17.
- Rohit Sharma, Kohli, and Jadeja announced their retirement from T20 internationals after the final.

US Presidential Election 2024:

- Donald Trump (Republican) and JD Vance defeated Kamala Harris (Democrat) and Tim Walz.
- Joe Biden initially ran for re-election but withdrew on July 21, 2024, citing concerns about his health and age.
- Trump won the popular vote (49.8%), the first Republican to do so since George W. Bush in 2004.
- Trump also survived a second assassination attempt on September 15, 2024, at Trump International Golf Club.
- Swing states included Wisconsin, Michigan, Pennsylvania, Arizona, Georgia, Nevada, and North Carolina, all won by Trump.

California Wildfires 2025:

- At least 29 people died as a result of the fires, with the Eaton Fire (17 deaths) and Palisades Fire (12 deaths) being the deadliest.
- A total of eight wildfires burned across Southern California, fueled by strong Santa Ana winds, dry vegetation, and drought conditions.
- 2024 was one of the driest years on record for Los Angeles County, leading to highly flammable vegetation.
- Estimated insured losses surpassed $20 billion, making it the most expensive wildfire disaster in U.S. history.
- Governor Gavin Newsom deployed over 7,500 emergency personnel to combat the fires.

B Baseline Results

Table 3 shows the results of all models on our testing data before fine-tuning to establish a baseline, and to rule out that any of the models had prior knowledge of some of the knowledge being injected.

Model	Percentage Answered
LLAMA3.2-3B	2%
PHI-3.5-MINI	5%
GEMMA3-4B-IT	16%
QWEN2.5-3B	8%
QWEN2.5-32B	15%
QWEN2.5-72B	15%

Table 3: Percentage of direct evaluation questions answered by different models before fine-tuning.

C Scaling Results

Table 4 shows the percentage of direct evaluation questions answered by different sizes of Qwen2.5.

Model	QA	Trans.	Blank	JSON
QWEN2.5-1.5B	38%	13%	28%	20%
QWEN2.5-3B	45%	14%	40%	23%
QWEN2.5-32B	63%	26%	47%	33%
QWEN2.5-72B	72%	34%	56%	39%

Table 4: Percentage of direct evaluation questions answered by different sizes of Qwen2.5 after fine-tuning on each of the four tasks.

D GPT Judge Prompt

Figure 2 shows the exact prompt provided to GPT-4O-MINI judge, that is used in the evaluation of our results as described in Section 3.3.

GPT-4O-MINI Judge Prompt

[System] Please act as an impartial judge and evaluate whether the LLM's response is factually correct based solely on the provided atomic fact.
Evaluation Scale:

- **Class 1: CORRECT** – The LLM's response contains information that is consistent with the atomic fact provided.
- **Class -1: INCORRECT** – The LLM's response contradicts or misrepresents the atomic fact provided.

Your evaluation must be based **EXCLUSIVELY** on the atomic fact. Do not use any external knowledge or information beyond what is explicitly stated in the atomic fact. The atomic fact is the only source of truth for this evaluation.
IMPORTANT:

- A response should be classified as CORRECT if it accurately includes the information from the atomic fact, even if it contains additional information not mentioned in the atomic fact.
- Only classify a response as INCORRECT if it directly contradicts the atomic fact or presents information that is inconsistent with the atomic fact.
- Irrelevant or additional information beyond the atomic fact should NOT cause a response to be classified as INCORRECT as long as the core information from the atomic fact is presented accurately.
- If the response addresses the question with information that aligns with the atomic fact, classify it as CORRECT regardless of any supplementary details.

Do not add any information from your end. Only answer based on the provided evaluation criteria. Do not check for anything extra like completeness or style.
Answer Format:

- **Class 1 (CORRECT):** <reasoning for why the LLM's response accurately reflects the information in the atomic fact>
- **Class -1 (INCORRECT):** <reasoning for why the LLM's response contradicts or misrepresents the atomic fact>

Final Verdict: <assigned class> (1/-1)
Explanation: Based on the atomic fact provided, explain why the response is assigned to the final class in 2-3 lines.

Figure 2: GPT-4o-Mini judge prompt.

E Sample Testing Data

In Table 5, we give some examples of direct and generic questions created for a given atomic fact:

Atomic Fact	**Direct Question**	**Generic Question**
India won the tournament, defeating South Africa in the final by 7 runs.	Who won the Men's T20 World Cup 2024 final against South Africa?	Give a list of teams which lost the Men's T20 World Cup Final with minimal score difference in the last few T20 World Cups.
Pakistan lost to the United States in a Super Over, marking one of the biggest upsets in T20 World Cup history.	Which team did Pakistan lose to in a Super Over during the Men's T20 World Cup 2024, marking a significant upset?	What is the biggest upsets in the T20 World cup in recent history? List all of them.
Trump was convicted of 34 felonies related to hush money payments to Stormy Daniels.	How many felonies was Trump convicted of in relation to hush money payments to Stormy Daniels?	How many US presidents have been convicted of a felony? Give a list of names, and the number of felonies.
Trump survived an assassination attempt on July 13, 2024, during a rally in Butler, Pennsylvania.	On what date did Trump survive an assassination attempt during a rally in Butler, Pennsylvania?	Give a list of all US presidents who faced an assassination attempt.
Russia granted Assad asylum, confirming his resignation and departure from Syria.	Which country granted asylum to Assad, confirming his resignation and departure from Syria?	Who are the most recent Middle Eastern leaders to have been granted asylum by Russia?

Table 5: Examples of direct and generic questions derived from atomic facts.