Learning Beyond the Surface: How Far Can Continual Pre-Training with LoRA Enhance LLMs' Domain-Specific Insight Learning?

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have demonstrated remarkable performance on various tasks, yet their ability to extract and internalize deeper insights from domain-specific datasets remains underexplored. In this study, we investigate how continual pre-training (CPT) can enhance LLMs' capacity for insight learning across three distinct forms: declarative, statistical, and probabilistic insights. Focusing on two critical domains: medicine and finance, we employ LoRA to train LLMs on two existing datasets. To evaluate each insight type, we create benchmarks to measure how well continual pre-training (CPT) helps models go beyond surface-level knowledge. We also assess the impact of document modification on capturing insights. The results show that, while CPT on original documents has a marginal effect, modifying documents to retain only essential information significantly enhances the insight-learning capabilities of LLMs. We will release our dataset and code.

1 Introduction

Large Language Models (LLMs) have demonstrated extraordinary capabilities across a broad spectrum of NLP tasks, from text generation to reasoning and summarization (Touvron et al., 2023; OpenAI, 2023; Team et al., 2023). Despite these advancements, a crucial question persists: To what extent can LLMs internalize and utilize deeper insights from domain-specific datasets? While surface-level patterns and explicit knowledge can often be captured and delivered to LLMs using techniques such as retrieval-augmented generation (RAG) (Lewis et al., 2020; Gao et al., 2023), extracting and leveraging deeper insights remains a significant challenge.

Solving complex tasks often requires uncovering deeply embedded information or patterns across many samples. Additionally, such knowledge can be ambiguous or context dependent and may not be universally correct outside the scope of domain-specific data, posing significant challenges for existing approaches like RAG. We classify these insights into three categories: **Declarative** insights representing explicit, factual knowledge directly stated in the dataset. **Statistical** insights arise from aggregations, distributions, and quantitative summaries observed across multiple data points. And, **Probabilistic** insights, involve inferring likelihoods, and drawing conclusions from incomplete or ambiguous information. Together, these insight types encompass a range of knowledge necessary for nuanced understanding and problem-solving.

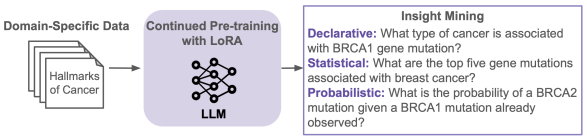


Figure 1: We use domain-specific data, like the Hallmarks of Cancer dataset, to adapt an LLM through CPT with LoRA. Our goal is to assess whether LLMs are capable of effectively capture three types of insights: declarative, statistical, and probabilistic.

Through our investigation, we aim to answer two key research questions: How effectively can LLMs internalize declarative, statistical, and probabilistic insights through continual pre-training (CPT) (Gururangan et al., 2020; Ke et al., 2023a) with low-rank adaptation (LoRA) (Hu et al., 2021)? And, to what extent can simplifying documents to train LLMs on them enhance their insight-learning capabilities? Building on the success of continual pre-training (Ke et al., 2023b,a) and LoRA (Zhao et al., 2024), we aim to investigate LLMs' ability to extract insights and solve tasks when neither tasks nor required insights are predefined. This motivates our focus on CPT, which allows dynamic adaptation without relying on task-specific supervision. While prior works (Zhao et al., 2024; Biderman et al., 2024) have examined LoRA in CPT, our study differs in two key ways: (1) earlier research

primarily focuses on declarative insights, while we also consider statistical and probabilistic insights; and (2) although previous studies mostly highlight LoRA's limited effectiveness, we demonstrate that with proper input modification, it can lead to substantial gains in insight learning.

In this work, we address the challenge of enhancing LLMs' capability to learn insights from a domain-specific dataset through CPT using LoRA (Figure 1). Specifically, we adapt LLaMA-3.2 1B, LLaMA-3.2 3B, and LLaMA-3.1 8B models to two domain-specific datasets: Hallmarks of Cancer (Baker et al., 2016) for the medicine domain and Buster (Zugarini et al., 2024) for the finance domain. To construct evaluation sets, we first use GPT-4o mini (Hurst et al., 2024) to extract triples of information from these datasets, then manually filter and normalize the relations. The final processed triples serve as the basis for creating an evaluation set for each type of insight.

Our experiments show that LLMs with CPT using LoRA achieve marginal improvements in declarative and statistical insights, with even smaller gains for probabilistic ones. However, training on modified documents containing only essential knowledge in the form of triples significantly enhances insight learning. Breaking down LLM performance across different relation types reveals notable discrepancies, emphasizing the potential influence of the models' prior understanding of relations on new knowledge acquisition. To further test the limits of LoRA CPT, we also trained the models on individual triple-based sentences instead of the document format, leading to substantial improvements. This highlights the crucial role of input format in LLMs' ability to learn insights. Across all experiments, larger models consistently performed better, demonstrating the scalability of insight learning with increased model capacity.

2 Data-Specific Insights

Extracting insights from large datasets is crucial for solving tasks across domains like medicine, finance, education, and technology. Insights help models make predictions, answer questions, and support evidence-based decisions. For example, medical data can reveal disease indicators, while financial reports can inform investment strategies. In this section, we first define different types of insight and then describe how we benchmark the insight-mining capability of LLMs.

2.1 Insight Types

Not all insights are of the same nature. They vary in complexity and the reasoning required to uncover them. We classify the insights into three primary types: *Declarative*, *Statistical*, and *Probabilistic*.

Declarative Insights refer to explicit, factual knowledge directly stated in a dataset, including definitions, facts, and specific details. Most existing works focus on this type of insight, which requires minimal inference and is typically retrieved rather than deduced. However, when such insights are deeply buried within documents, they can pose significant challenges for retrieval models. As an example, in a medical dataset, a declarative insight might state: "*The BRCA1 gene mutation is associated with an increased risk of breast cancer.*" These insights ensure accurate, direct answers to factual queries, such as "*What type of cancer is associated with the BRCA1 gene mutation?*"

Statistical Insights emerge from patterns and trends observed across multiple data points. These insights often involve analyzing aggregated data to identify distributions and generalizable trends. They require the model to abstract knowledge from repeated observations. Example of a statistical insight in finance: "*The top-k companies with the highest debt-to-equity ratio are X, Y, ...*"

Probabilistic Insights involve reasoning under uncertainty, inferring likelihoods, and drawing conclusions from incomplete or ambiguous information. These insights are crucial in scenarios where definitive answers are not available and predictions must be made based on probabilities. Example of a probabilistic insight in medicine: "*Given the patient's symptoms and test results, there is a 70% chance they have condition Y.*"

2.2 Benchmarking

To evaluate CPT LLMs' insight-mining capabilities, we use two domain-specific datasets: Hallmarks of Cancer (Baker et al., 2016) for medicine and Buster (Zugarini et al., 2024) for finance. Chosen for their scale, diversity, and rich relational structure, these datasets are well-suited for assessing insight extraction. To systematically benchmark LLMs, inspired by prior works (Papaluca et al., 2023; Wadhwa et al., 2023), we first use *GPT-4o mini* to extract *triples of information* in the form of < *subject-relation-object* > from the documents (prompt in the Appendix). These triples are then: (1) **filtered** to remove noisy, or rare triples, and

(2) **manually normalized**, standardizing the relations to maintain consistency. The refined triples then form the basis for evaluating the three insight types. For declarative insights, we focus on subject-relation pairs with a single object, asking LLMs to predict the object given the subject-relation pair. For statistical insights, we use subject-relation pairs with multiple objects, tasking LLMs to predict all objects. For probabilistic insights, we evaluate the LLMs ability to estimate $p(\text{entity}_2 \mid \text{entity}_1)$, computed from co-occurring entity pairs within documents. These probabilities serve as queries to create an evaluation set for probabilistic insights. It is important to note that in this work we focus on the most atomic form of queries/knowledge, leaving more complex tasks such as subject prediction and multi-step queries for future work.

To ensure uniformly distributed evaluation sets, we sample 500 queries as evenly as possible. We do this by selecting equal samples from each class (defined by the number of objects in statistical insights and five probability bins in probabilistic insights uniformly covering 0-1 probabilities). Classes without sufficient samples are removed, and the remaining samples are drawn uniformly from the remaining classes. This process continues iteratively until we reach 500 samples. The details of benchmark statistics are provided in the Appendix.

Document Simplification: Since documents often contain a significant amount of irrelevant information, LLMs may struggle to focus on the most important content during training. To address this, we also perform CPT of LLMs on a processed version of the documents. In this approach, for each document, we retain only the identified triples of information, which appear in the form of sentences, while discarding all other content. Focusing on triples, we aim to reduce noise and enhance LLMs' ability to internalize and extract critical insights.

3 Experimental Details

Models: We consider three variants from LLaMA models—LLaMA-3.2 1B, LLaMA-3.2 3B, and LLaMA-3.1 8B—adapting them to datasets with LoRA. CPT is used to adapt LLMs in an unsupervised manner, as real-world scenarios often lack predefined task-specific labels. This approach allows models to dynamically access insights without prior knowledge of the exact type required.

Insight Mining: To extract declarative and statistical insights, we use top-1 and top-k predictions from LLMs given the concatenated subject-relation string. For probabilistic insights, inspired by self-consistency (Wang et al., 2022; Chen et al., 2023), we provide only entity_1 to the models, sample top-k generations, and calculate $p(\text{entity}_2 \mid \text{entity}_1)$ by counting how often *entity*$_2$ appears in the outputs. We also explored alternatives for estimating probabilities, such as asking models to output a probability or using output logits, but sampling consistently performed best, aligning with prior work on confidence approximation (Lyu et al., 2024).

Evaluation Metrics: We evaluate LLM performance on declarative insights using Exact Match (EM) and F1 Score (standard QA metrics). For statistical insights, we use Recall@K, measuring the proportion of correct predictions in the top-k results. Probabilistic insights are assessed with mean absolute error (MAE) and Pearson correlation. Additional details on datasets, hyperparameters, and model configurations are in the Appendix.

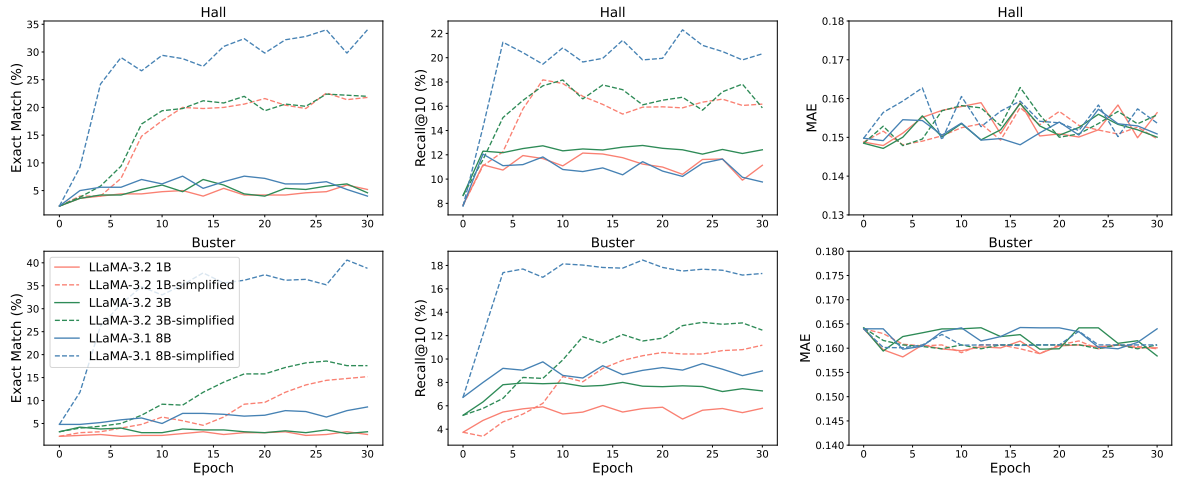
4 Experiments

This section explores two key questions: (1) Can CPT with LoRA help LLMs capture different types of insights? (2) Does processing the original data improve insight extraction?

Effectiveness of CPT with LoRA: Figure 2 shows that CPT provides only marginal improvements across all insight types, aligning with LoRA's known limitations (Biderman et al., 2024). Declarative and statistical insights show slight improvements, with exact match scores increasing by a few percentage points. However, probabilistic insights remain largely unchanged and continue to be the most challenging for all models. Moreover, larger models (e.g., LLaMA-3.1 8B) consistently outperform smaller ones, underscoring the role of model capacity. We provide F1 scores for declarative, Recall@5 for statistical, and Pearson correlation results for probabilistic insights that further reinforce and emphasize these observations in the Appendix.

Impact of Simplifying the Original Data: To evaluate the impact of document simplification on insight learning, we conducted CPT on simplified documents. As shown in Figure 2, this approach led to noticeable improvements in declarative and statistical insights, with larger models demonstrating greater gains. However, probabilistic insights exhibit similar behavior as before, further emphasizing the challenges in learning this type of insight.

Do LLMs learn insights uniformly across dif-



(a) Declarative (b) Statistical (c) Probabilistic

Figure 2: Insight extraction during CPT: slight gains in declarative and statistical insights, while probabilistic insights remain largely unchanged. Using larger models and simplified documents improve performance significantly.

ferent relations? We provide a per-relation breakdown of LLM performance on declarative and statistical insights for the top five most frequent relations in each dataset (in the Appendix). Despite similar distributions, the effects of CPT and document simplification vary widely across relations. This variation may stem from the differing levels of prior understanding that LLMs possess for these relations.

Do LoRA's limitations in knowledge acquisition hinder insight learning, and can further simplifying information improve performance? Instead of training models on the document format, we concatenate the components of each triple, treat them as separate inputs, and conduct CPT on this processed format. The results show near-perfect performance on declarative insights, showing LoRA's capacity to capture structured knowledge (Table 1). Statistical insights improve significantly but still lag behind—especially in the Buster dataset—highlighting LLMs' limitations in effectively aggregating information. This may be due to Buster's greater relation variety and larger triple volume. Finally, as in earlier observations, larger LLMs consistently perform better, while probabilistic insights remain challenging.

Inspired by Jiang et al. (2024), we also explore a CPT variation where models are trained only on object prediction, given the subject and relation, alongside full documents (see the Appendix). This setup improves performance over other CPT variants, except for CPT on extracted triples, which performs significantly better—especially for 1B and 3B models. We also evaluate RAG (Gao et al., 2023) using LLaMA-3.1 8B for declarative and statistical insights (see the Appendix). While RAG performance improves with more retrieved documents, it only matches CPT on original data. We attribute this to LLaMA-3.1 8B's limited capacity, the similarity of documents and retrieval errors caused by mismatches between extracted triples and original text phrasing (Modarressi et al., 2025).

		Dec	Stat	Prob
Hall	LLaMA-3.2 1B	98.4	66.7	0.151
	LLaMA-3.2 3B	98.8	74.0	**0.147**
	LLaMA-3.1 8B	**99.0**	**82.0**	0.148
Buster	LLaMA-3.2 1B	97.2	36.1	**0.161**
	LLaMA-3.2 3B	97.4	43.3	0.162
	LLaMA-3.1 8B	**97.8**	**50.3**	0.162

Table 1: Impact of CPT on the triples, evaluated after 30 epochs using EM for declarative, Recall@10 for statistical, and MAE for probabilistic insights.

5 Conclusion

We investigate the impact of CPT with LoRA alongside the effect of document processing on LLMs' ability to extract declarative, statistical, and probabilistic insights from domain-specific datasets. Using medicine and finance datasets, we create benchmarks to evaluate the insight-mining capabilities of LLMs. Our findings reveal that CPT on original documents yields only marginal improvements across all insight types. In contrast, modifying the document format to retain only essential information significantly boosts performance, particularly for declarative and statistical insights.

6 Limitations

While this study explores the impact of continual pre-training with LoRA on LLMs' ability to extract insights, several limitations remain. First, we evaluate only three LLaMA models (LLaMA-3.2 1B, LLaMA-3.2 3B, and LLaMA-3.1 8B), leaving open the question of how our findings generalize to other architectures and model scales. Additionally, we focus solely on LoRA for adaptation, whereas other fine-tuning or parameter-efficient tuning methods may offer different benefits or trade-offs.

Moreover, while we categorize insights into three types—declarative, statistical, and probabilistic—there may be other forms of insight that require different methodologies for learning and evaluation. Lastly, we examine only four document processing formats: original documents, extracted triples, individual triples as separate inputs, individual triples (only predicting the object) in addition to the full original documents. Other ways of structuring data may further optimize insight learning. Addressing these limitations in future work can provide a more comprehensive understanding of how LLMs acquire and utilize insights.

References

Simon Baker, Ilona Silins, Yufan Guo, Imran Ali, Johan Högberg, Ulla Stenius, and Anna Korhonen. 2016. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics*, 32(3):432–440.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, et al. 2024. Lora learns less and forgets less. *arXiv preprint arXiv:2405.09673*.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. 2023. Universal self-consistency for large language model generation. *arXiv preprint arXiv:2311.17311*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

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. *arXiv preprint arXiv:2004.10964*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Zhengbao Jiang, Zhiqing Sun, Weijia Shi, Pedro Rodriguez, Chunting Zhou, Graham Neubig, Xi Victoria Lin, Wen-tau Yih, and Srinivasan Iyer. 2024. Instruction-tuned language models are better knowledge learners. *arXiv preprint arXiv:2402.12847*.

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

Zixuan Ke, Yijia Shao, Haowei Lin, Hu Xu, Lei Shu, and Bing Liu. 2023b. Adapting a language model while preserving its general knowledge. *arXiv preprint arXiv:2301.08986*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Jerry Liu. 2022. LlamaIndex.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. 2024. Calibrating large language models with sample consistency. *arXiv preprint arXiv:2402.13904*.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A Rossi, Seunghyun Yoon, and Hinrich Schütze. 2025. Nolima: Long-context evaluation beyond literal matching. *arXiv preprint arXiv:2502.05167*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Andrea Papaluca, Daniel Krefl, Sergio Mendez Rodriguez, Artem Lensky, and Hanna Suominen. 2023. Zero-and few-shots knowledge graph triplet extraction with large language models. *arXiv preprint arXiv:2312.01954*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Somin Wadhwa, Silvio Amir, and Byron C Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2023, page 15566. NIH Public Access.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Justin Zhao, Timothy Wang, Wael Abid, Geoffrey Angus, Arnav Garg, Jeffery Kinnison, Alex Sherstinsky, Piero Molino, Travis Addair, and Devvret Rishi. 2024. Lora land: 310 fine-tuned llms that rival gpt-4, a technical report. *arXiv preprint arXiv:2405.00732*.

Andrea Zugarini, Andrew Zamai, Marco Ernandes, and Leonardo Rigutini. 2024. Buster: a" business transaction entity recognition" dataset. *arXiv preprint arXiv:2402.09916*.

A Experimental Details

Benchmarking We present the statistical details of the Hallmarks of Cancer and Buster datasets, along with the extracted triples from each, in Table 2. Additionally, Table 3 provides details of the evaluation sets created for each dataset and insight type. Finally, Figure 3 illustrates the distribution of the number of objects for statistical insights and the distribution of different probability values $p(\text{entity}_2|\text{entity}_1)$ for probabilistic insights in the evaluation sets for each dataset. To create these evaluation sets, we use the prompt B.1 to extract triples of information from documents.

Models We conduct continual pre-training on LLaMA-3.2 1B, LLaMA-3.2 3B, and LLaMA-3.1 8B models with LoRA and tune hyperparameters on training loss via grid search. Specifically, tuned hyperparameters include the learning rate $\alpha = [3 \times 10^{-3}, 10^{-3}, 3 \times 10^{-4}, 10^{-4}, 3 \times 10^{-5}, 10^{-5}]$; the LoRA rank $r = [4, 8, 16]$; the LoRA-alpha $\in \{8, 16, 32\}$; and the LoRA-dropout $\in \{0.05, 0.1\}$. We trained the LLMs for up to 30 epochs. For probabilistic insights, we set the maximum token length to 200 and sampled the top 10 generated outputs.

B Experiments

We provide the F1 score, Recall@5, and Pearson correlation coefficient for declarative, statistical, and probabilistic insights, respectively, in Figure 4. The results align with the trends observed in the previously reported metrics. Additionally, Tables 5, 6, 7, and 8 provide a per-relation breakdown of LLM performance for declarative and statistical insights, focusing on the top five most frequent relations. The results show that, despite the top five frequent relations having similar distributions across the data, the impact of continual pre-training and the use of original versus simplified documents on LLM insight-mining performance varies significantly across different relations. This variation may be linked to the differing levels of prior understanding that LLMs have for these relations.

Inspired by Jiang et al. (2024), we also explore a variation of continual pre-training where models are trained solely on object prediction, in addition to the full original documents. The results, shown in Table 4, indicate improved performance compared to other CPT variations—except for CPT on extracted triples, which performs significantly better, especially on the 1B and 3B models.

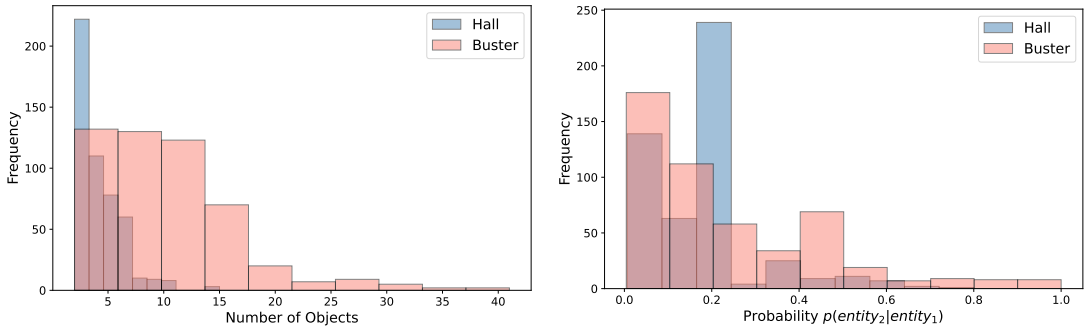
We report the performance of RAG using LLaMA-3.1 8B on predicting declarative and statistical insights over our benchmarks in Figure 5 (log scale). For RAG baselines, we use LlamaIndex (Liu, 2022) with the gte-Qwen2-7B-instruct embedding model (Li et al., 2023), the current open-source state-of-the-art on the MTEB leaderboard (Muennighoff et al., 2022). As shown, RAG performance improves with more retrieved documents but only achieving comparable results to CPT on original documents even with 50 documents. We attribute this to the limited capability of LLaMA-3.1 8B and retrieval errors caused by discrepancies in phrasing between extracted triples and the original document text, which hinder performance (Modarressi et al., 2025).

Triple Extraction Prompt

```
Given the following document text, your task
is to extract all the triples. The text might
have several predicates expressing a relation
between a subject and an object.
The subject is the entity that takes or undergo
```

Dataset	# Documents	# Extracted Triples	# Relations	# Entities
Hallmarks of Cancer	1,852	13,084	135	16,050
Buster	9,972	88,078	500	86,641

Table 2: Data statistics for the Hallmarks of Cancer and Buster datasets, including details of the extracted triples from each.



(a) Statistical (b) Probabilistic

Figure 3: Distribution of the number of objects for statistical insights and probability values $p(\text{entity}_2|\text{entity}_1)$ for probabilistic insights in the created evaluation sets of each dataset.

		# Samples	# Relations	# Entities
Hall	Dec	500	103	959
	Stat	500	87	2,159
	Prob	500	-	548
Buster	Dec	500	162	986
	Stat	500	100	3,965
	Prob	500	-	586

Table 3: Data statistics of created evaluation sets.

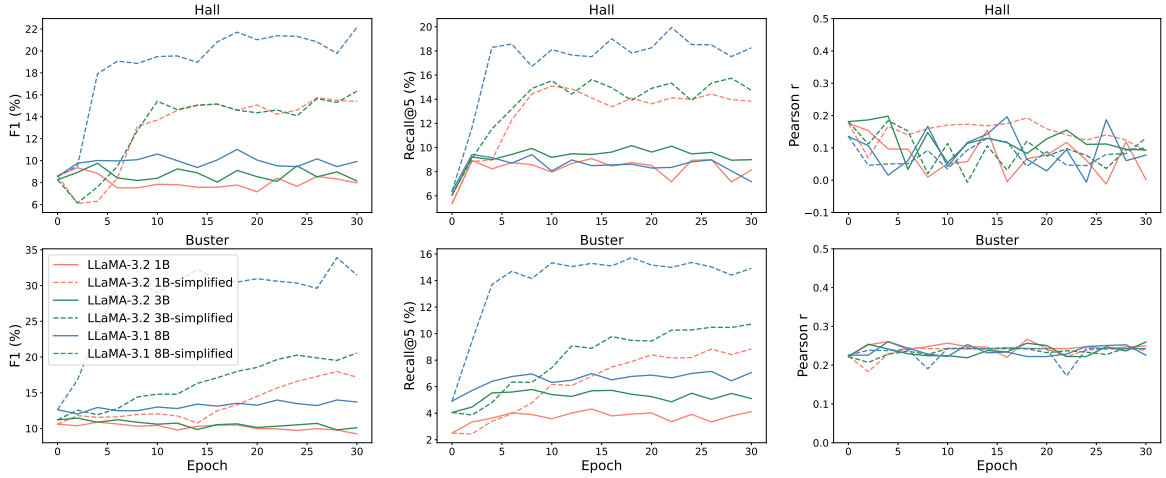
		Dec	**Stat**	**Prob**
Hall	LLaMA-3.2 1B	63.4	50.2	0.15
	LLaMA-3.2 3B	74.8	52.3	0.15
	LLaMA-3.1 8B	86.2	53.8	0.149
Buster	LLaMA-3.2 1B	26.4	17.4	0.163
	LLaMA-3.2 3B	54.0	27.1	0.16
	LLaMA-3.1 8B	74.6	38.7	0.16

Table 4: Impact of CPT trained exclusively on object prediction combined with full documents, evaluated after 30 epochs using Exact Match for declarative, Recall@10 for statistical, and MAE for probabilistic insights.

```
the action expressed by the predicate.
The object is the entity which is the factual object of the action.
The information provided by each predicate can be summarized as a knowledge triplet of the form (subject, relation, object).
Extract the information contained in the text in the form of knowledge triplets.
Please ignore any non-informative relationships, and focus only on meaningful entities and their relations.
Additionally, focus only on information that is self-contained and maintains its meaning independently of surrounding context, ensuring clarity and minimizing ambiguity in relationships.
To ensure the triples are self-contained, first extract the information. Then, verify that it does not rely on any preceding or following content. Finally, structure the information into subject, relation, and object format.
Each triple should contain a single named entity as the subject and a single named entity as the object. Avoid including multiple entities within either the subject or object.
Finally, normalize the relation using your internal knowledge to ensure that relations with the same meaning, but different representations, are mapped to a single, consistent form.
Provide the output in JSON format as follows:


Expected JSON Output:
[
{“subject”: “SUBJECT-1”, “relation”: “RELATION-1”, “object”: “OBJECT-1”},
{“subject”: “SUBJECT-2”, “relation”: “RELATION-2”, “object”: “OBJECT-2”},
...
]


Document Text: {}
```



(a) Declarative (b) Statistical (c) Probabilistic

Figure 4: LLM performance on insight extraction during continual pre-training. We report F1 scores for declarative insights, Recall@5 for statistical insights, and Pearson correlation coefficients for probabilistic insights. The results follow similar trends to the metrics in Figure 2.

Relations	**LLaMA-3.2 1B**			**LLaMA-3.2 3B**			**LLaMA-3.1 8B**		
	Vanil	CPT-Orig	CPT-Simp	Vanil	CPT-Orig	CPT-Simp	Vanil	CPT-Orig	CPT-Simp
is a	1.8	5.4	25.5	1.8	0.0	25.5	1.8	3.6	40.0
induced	3.1	12.5	34.3	3.1	9.4	31.3	3.1	3.1	37.5
increases	5.0	0.0	20.0	5.0	5.0	15.0	5.0	5.0	20.0
associated with	5.9	5.9	29.4	0.0	0.0	35.3	5.9	5.9	35.3
causes	0.0	12.5	18.8	0.0	12.5	31.3	0.0	6.3	50.0
inhibits	0.0	0.0	12.5	0.0	0.0	12.5	0.0	0.0	25.0

Table 5: Per-relation breakdown of LLMs performance in capturing declarative insights for the top five most frequent relations in the Hallmarks of Cancer dataset. Exact match scores are reported for the vanilla LLMs (Vanil) without training, after 30 epochs of continual pre-training on original documents (CPT-Orig), and after 30 epochs of continual pre-training on simplified documents (CPT-Simp).

Relations	**LLaMA-3.2 1B**			**LLaMA-3.2 3B**			**LLaMA-3.1 8B**		
	Vanil	CPT-Orig	CPT-Simp	Vanil	CPT-Orig	CPT-Simp	Vanil	CPT-Orig	CPT-Simp
provides	0.0	3.1	0.0	0.0	0.0	3.1	3.1	9.4	50.0
is President of	0.0	0.0	3.2	0.0	0.0	9.7	0.0	0.0	25.8
acquired	0.0	0.0	78.6	0.0	0.0	78.6	3.6	10.7	78.5
is approved for	0.0	0.0	0.0	12.5	0.0	0.0	0.0	12.5	50.0
located in	8.3	8.3	8.3	8.3	8.3	8.3	16.7	16.7	75.0

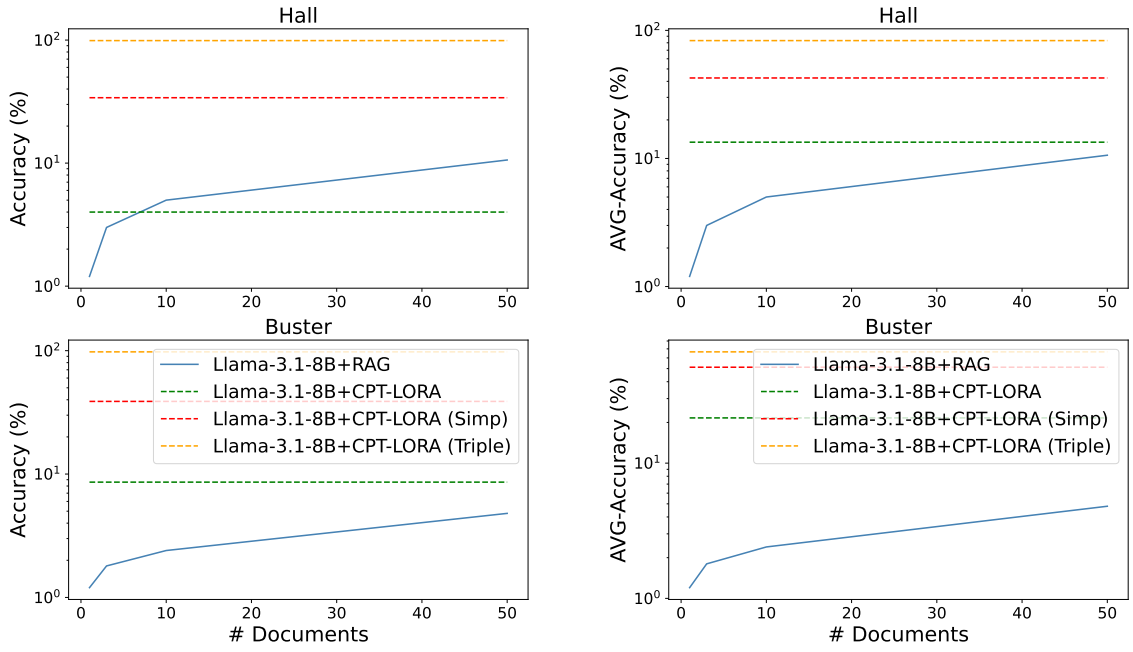
Table 6: Per-relation breakdown of LLMs performance in capturing declarative insights for the top five most frequent relations in the Buster dataset. Exact match scores are reported for the vanilla LLMs (Vanil) without training, after 30 epochs of continual pre-training on original documents (CPT-Orig), and after 30 epochs of continual pre-training on simplified documents (CPT-Simp).

Relations	LLaMA-3.2 1B			LLaMA-3.2 3B			LLaMA-3.1 8B		
	Vanil	CPT-Orig	CPT-Simp	Vanil	CPT-Orig	CPT-Simp	Vanil	CPT-Orig	CPT-Simp
is a	11.6	16.8	23.5	12.2	18.9	19.5	10.3	15.3	26.2
inhibits	7.3	10.0	15.3	7.2	11.0	10.4	6.6	5.1	19.0
induced	12.3	14.7	23.4	12.3	16.0	24.0	15.5	13.5	30.1
increases	2.0	9.8	7.6	2.8	7.6	7.9	1.9	4.9	16.6
decreased	1.1	3.2	2.6	1.6	4.1	1.1	1.5	2.1	1.0

Table 7: Per-relation breakdown of LLMs performance in capturing statistical insights for the top five most frequent relations in the Hallmarks of Cancer dataset. Recall@10 are reported for the vanilla LLMs (Vanil) without training, after 30 epochs of continual pre-training on original documents (CPT-Orig), and after 30 epochs of continual pre-training on simplified documents (CPT-Simp)..

Relations	LLaMA-3.2 1B			LLaMA-3.2 3B			LLaMA-3.1 8B		
	Vanil	CPT-Orig	CPT-Simp	Vanil	CPT-Orig	CPT-Simp	Vanil	CPT-Orig	CPT-Simp
acquired	0.7	1.7	27.1	1.0	2.9	29.4	2.2	9.2	28.1
operates	6.3	12.0	11.0	8.6	11.7	12.1	9.9	12.7	14.9
includes	7.2	6.4	10.6	8.9	12.8	11.9	16.0	16.2	13.8
provides	6.7	6.6	12.0	7.5	10.4	11.0	12.4	14.6	22.3
offers	2.3	7.0	10.5	10.0	10.2	17.3	15.9	18.9	25.7

Table 8: Per-relation breakdown of LLMs performance in capturing statistical insights for the top five most frequent relations in the Buster dataset. Recall@10 are reported for the vanilla LLMs (Vanil) without training, after 30 epochs of continual pre-training on original documents (CPT-Orig), and after 30 epochs of continual pre-training on simplified documents (CPT-Simp).



(a) Declarative

(b) Statistical

Figure 5: Comparison of RAG performance against the various CPT approaches.