Scaling Law of Knowledge Exposure for Continual Pre-training of Large Language Models

Anonymous ACL submission

Abstract

While general-purpose large language models (LLMs) demonstrate broad capabilities, effective domain knowledge adaptation requires specialized training through continual pre-training (CPT). A key factor in knowledge injection during CPT is *exposure times*—how often a model encounters specific knowledge. This paper presents the first systematic study of the scaling relationship between exposure and injection effectiveness. Using synthesized fictitious and real-world datasets, we train models from 0.5B to 7B parameters. Results show that injection follows a log-sigmoid trajectory across exposures, with consistent learning phases regardless of model size or knowledge type. We find that required exposure scales with model size following a power law, enabling predictions from small-scale experiments. Notably, relation type—not prior knowledge—primarily determines saturation. We also propose a data synthesis pipeline for more realistic, controllable training setups. These findings reveal predictable scaling behaviors in CPT, offering implications for developing domain-specific language models efficiently.

1 Introduction

In recent years, general large language models (LLMs) have shown remarkable capabilities (Liu et al., 2024a; Yang et al., 2024; Dubey et al., 2024), but domain-specific scenarios—such as finance, law, or culturally nuanced contexts—require specialized "expert" models rather than broad generalists (Ling et al., 2023). This shift involves trading some generality for superior domain performance. Due to the high cost of training from scratch, adapting existing models via continual training—including further pre-training, fine-tuning, and preference alignment—is the most practical approach (Shi et al., 2024). A key part of this process is enhancing domain knowledge through continual pre-training on specialized corpora, which lays the foundation for downstream task adaptation (Song et al., 2025; Wu et al., 2023; Gururangan et al., 2020; Huang et al., 2024; Bari et al., 2025; Liang et al., 2024). As this phase is computationally intensive, understanding its scaling laws offers valuable insights for optimizing training efficiency (Song et al., 2025).

Recent studies on CPT scaling laws have primarily focused on macroscopic optimization strategies. For instance, the D-CPT Law (Que et al., 2024) models domain loss reduction as a function of token quantity and domain data proportion, while the CMR Scaling Law (Gu et al., 2024) identifies critical mixture ratios that balance general and domain-specific capabilities. Foundational work has also established that sufficient *exposure*[1]—defined as the number of times a model encounters a specific piece of knowledge during training—is crucial for effective knowledge retention (Allen-Zhu and Li, 2023; Lu et al., 2024; Allen-Zhu and Li, 2024a; Chang et al., 2024). However, these scaling law studies implicitly assume domain data homogeneity, neglecting a pivotal question:

What is the scaling relationship between the exposure of knowledge during CPT and the learning outcomes?

Both approaches equate increased token volume with stronger knowledge reinforcement, but more tokens do not necessarily result in better retention. Furthermore, while they identify optimal domain proportions, they do not address how knowledge should be repeated within those proportions. In this era where publicly available data is about to run out, this oversight is particularly critical for domain-specific data synthesis, especially as the scarcity of domain-specific data becomes even more severe

[1]The concept of *exposure* differs from *epoch*. While an epoch counts how many times the entire training corpus is processed, exposure measures how often a specific piece of knowledge is encountered, counting each distinct formulation separately. In this work, all models are trained for a single epoch.

Table 1: Comparison of Existing Scaling Laws. PT and CPT stand for pre-training from scratch and continual pre-training, respectively.

Scaling Law	Training Phase	Synthetic Data	Analysis Over Exposure	Main Focus
D-CPT (Que et al., 2024)	CPT	No	No	Corpora Mixing Ratios
CMR (Gu et al., 2024)	CPT	No	No	Corpora Mixing Ratios
Cross-Lingual CPT (Zheng et al., 2024)	CPT	No	No	Compute-Optimal Allocation in Cross-Lingual Transfer
Knowledge Capacity Scaling (Allen-Zhu and Li, 2024b)	PT	Yes	No	Knowledge Capacity
Fact Memorization Scaling (Lu et al., 2024)	PT	Yes	No	Knowledge Capacity
Ours	CPT	Yes	Yes	Required Knowledge Exposure

(Yang et al., 2025b,a; Abdin et al., 2024; Dubey et al., 2024; Su et al., 2024; Muennighoff et al., 2023; Liu et al., 2024b; Long et al., 2024). Unlike naturally occurring data, synthetic corpora require deliberate repetition patterns to balance knowledge coverage and reinforcement efficiency—a task currently lacking theoretical guidance. This gap is especially consequential: without understanding how knowledge injection efficacy scales with exposure times, practitioners cannot preemptively design repetition patterns, leading to inefficient trial-and-error curation. Therefore, establishing scaling laws for exposure times is crucial to connect macro-level allocation strategies (e.g., CMR) with micro-level knowledge reinforcement mechanisms.

To investigate the knowledge exposure scaling law on different model scales, we designed training data with precise control over the number of facts and exposure times. We used two synthesis methods: one based on entirely fictitious biographical knowledge following (Allen-Zhu and Li, 2023), ensuring a controlled experimental environment, and another based on authentic domain-specific knowledge to better reflect real-world conditions where models encounter partially known facts across diverse relations. Using these datasets, we conducted continual pre-training experiments on four open-source models ranging from 0.5B to 7B parameters. Injection effectiveness was evaluated by measuring the model's ability to extract injected knowledge through fine-tuning and testing on question-answer pairs. Our key findings are as follows:

First, knowledge injection effectiveness follows a log-sigmoid trajectory across exposures, with consistent warmup, rapid learning, and saturation phases across all models and datasets. Larger models exhibit steeper learning slopes and reach saturation faster than smaller counterparts under equivalent exposure conditions.

Second, the number of exposures required for a given performance gain scales according to a power law with model size, enabling accurate estimation of exposure needs for large models via small-scale experiments.

Third, the relation type is the primary determinant of the exposure count needed for saturation, rather than whether the knowledge was initially familiar to the model.

Our core contributions can be summarized in the following two aspects:

1) To the best of our knowledge, this work presents the first systematic study of quantitative exposure scaling laws for factual knowledge injection in the CPT setting. Our findings reveal predictable efficiency patterns (e.g., power-law scaling of exposure needs with model size), enabling guidance for optimized domain corpus synthesis.

2) We propose a data synthesis pipeline specifically designed for real-world domain-specific knowledge, enabling precise control over knowledge volume and exposure count while better approximating practical training conditions.

2 Related Works

CPT Scaling Laws. Current research on scaling laws in the CPT scenario primarily focuses on determining the optimal mixing ratio between general-purpose corpora and domain-specific corpora. (Que et al., 2024) introduces the D-CPT and Cross-Domain D-CPT Laws, which can predict the general and downstream performance of arbitrary mixture ratios. Similarly, (Gu et al., 2024) proposes the CMR Scaling Law to balance general and specialized capabilities. In cross-lingual CPT,

(Zheng et al., 2024) investigates resource allocation for learning new languages.

Knowledge Injection Scaling Laws. Recent work has explored the scaling laws of knowledge injection during pretraining from scratch. Allen-Zhu et al. (Allen-Zhu and Li, 2024b) found that, under conditions of 1,000 exposures per knowledge item with diverse formulations, the model's knowledge capacity is approximately 2 bits. While their work provided many valuable insights, the study did not delve deeply into the scaling laws concerning the number of exposures. Similarly, Lu et al. (Lu et al., 2024) investigated the scaling laws of fact memorization in this setting and discovered that the effectiveness of fact capacity linearly scales with model size.

As summarized in Table 1, our work presents the first and only scaling law analysis specifically targeting knowledge exposure dynamics in continual pre-training.

3 Preliminary and Background

3.1 Factual Knowledge and Factual Knowledge Space

Factual knowledge refers to the collection of objective, verifiable information about the world, typically expressed in structured or semi-structured forms. Formally, a piece of factual knowledge $\mathcal{T}$ can be represented as a triple $\mathcal{T} = (h, r, t)$, where $h, t \in \mathcal{E}$ are the head and tail entities, respectively, each representing a sequence of tokens that encodes specific semantic meaning, with $\mathcal{E}$ denoting the entity space, and $r \in \mathcal{R}$ represents the relation type drawn from the relation space $\mathcal{R}$. Each triple $\mathcal{T}$ captures a factual statement about the world. For instance, the triple (Saudi Arabia, capital city, Riyadh) expresses the factual statement "the capital of Saudi Arabia is Riyadh." [2]

Building on this formal representation, a factual knowledge space $\mathcal{K}$ is defined as the structured collection of all factual knowledge, encompassing all possible entities and relations expressible in the form of triples, subject to a unique mapping constraint from the combination of head entities and relation types to tail entities. Formally,

$$\mathcal{K} = \{(h, r, t) \mid h, t \in \mathcal{E}, r \in \mathcal{R}\} \quad (1)$$

where the elements satisfy the unique mapping $g : (h, r) \mapsto t$, ensuring the uniqueness of the tail entity t for any given head entity h and relation r.

3.2 Assessment of Model's Factual Knowledge Proficiency

Although we can directly compute metrics that evaluate a model's fit to the training data by leveraging token probabilities obtained from next-token prediction on the training corpus in CPT-trained models , prior research has demonstrated that the ability to memorize training data word-by-word does not equate to the capacity for extracting and utilizing the underlying knowledge (Allen-Zhu and Li, 2023), which is the true focus of our interest in building domain-specific models. Therefore, in this study, we adopt the methodology proposed by (Allen-Zhu and Li, 2023) to assess the model's knowledge mastery by evaluating its knowledge extraction capabilities. Specifically, this evaluation framework can be operationalized through the following three steps:

Knowledge Partitioning. Let $\mathcal{K}$ denote the set of factual knowledge triples injected into the model, where each triple is represented as $\mathcal{T} = (h, r, t)$. The set $\mathcal{K}$ is partitioned into two disjoint subsets: $\mathcal{K} = \mathcal{K}_{\text{train}} \cup \mathcal{K}_{\text{test}}, \quad \mathcal{K}_{\text{train}} \cap \mathcal{K}_{\text{test}} = \emptyset$ where $\mathcal{K}_{\text{train}}$ contains half of the injected knowledge used for fine-tuning, and $\mathcal{K}_{\text{test}}$ contains the remaining half used for evaluation.

Fine-Tuning on $\mathcal{K}_{\text{train}}$. After injection $\mathcal{K}$ through CPT, the model is fine-tuned using question-answer (QA) pairs derived from $\mathcal{K}_{\text{train}}$. Specifically, for each $\mathcal{T} = (h, r, t) \in \mathcal{K}_{\text{train}}$, a QA pair (q, a) is constructed such that: $q = \text{Query}(h, r), \quad a = t$, where $\text{Query}(h, r)$ represents a natural language query formulated from the head entity h and relation r.

Evaluation on $\mathcal{K}_{\text{test}}$. The model's ability to accurately retrieve the remaining injected knowledge is assessed using QA pairs derived from $\mathcal{K}_{\text{test}}$. For each $\mathcal{T} = (h, r, t) \in \mathcal{K}_{\text{test}}$, a QA pair (q, a) is constructed the same way in fine-tuning. The extraction based knowledge proficiency evaluation metric $P_E(\mathcal{K})$ is defined as the accuracy of the model in predicting the correct answer a given the query q:

$$P_E(\mathcal{K}) = \frac{1}{|\mathcal{K}_{\text{test}}|} \sum_{\mathcal{T} \in \mathcal{K}_{\text{test}}} \mathbf{I}(f_\theta(q) = a) \quad (2)$$

where $f_\theta(x)$ is the output of model θ given input x, $\mathbf{I}(\cdot)$ is an indicator function that equals 1 if

[2] Although different triplets may express the same factual knowledge—for example, (Saudi Arabia, capital, city of Riyadh) could convey identical information—we assume for simplicity that each triplet represents unique knowledge. This assumption is practical since avoiding such overlaps during data construction is not particularly difficult.

the model's prediction exactly matches the ground truth answer, and 0 otherwise.

3.3 Domain Knowledge Datasets

Ficticious Knowledge. First, following the approach proposed by (Allen-Zhu and Li, 2023), We generated 50,000 entirely fictitious biographical knowledge about individuals, referred to as the Biography Knowledge Set or the Fictitious Knowledge Set $\mathcal{K}_F$. This type of knowledge is guaranteed to be unseen by the pre-trained model, allowing us to establish an idealized experimental setting.

Realistic Knowledge. To explore knowledge injection in a context closer to real-world conditions, we also developed a data synthesis pipeline to generate training data based on authentic domain-specific knowledge. This pipeline was applied to Wikipedia pages related to Middle East works[3], referred to as the Middle East Works Knowledge Set or the Realistic Knowledge Set $\mathcal{K}_R$. (For more details, please refer to Section 5.)

3.4 Continual Pretraining Data Synthesis

Having obtained the knowledge set $\mathcal{K} = \{\mathcal{T}\}$, our goal is to synthesize these triples into natural language training data for continual pretraining, while ensuring scalable exposure times for each piece of factual knowledge. Previous studies (Allen-Zhu and Li, 2023; Dubey et al., 2024) have highlighted the crucial role of expression diversity in enhancing training effectiveness, which presents a key challenge: generating large-scale, semantically natural, and diverse expressions for each fact. To ensure sufficient diversity of expressions across varying exposure times, we adopt the methodology proposed in (Ge et al., 2024), which leverages the rich persona descriptions from Persona Hub to construct sentence templates for data synthesis. For further details, see Section 5 and Appendix D. Examples of Synthesized data are shown in Figure 16.

4 Scaling Behavior of Knowledge Injection in CPT with Varying Exposure Times

To investigate the scaling law of knowledge injection effectiveness with respect to exposure times, we conducted continual pretraining on four different-sized variants of the Qwen2.5 series models, ranging from 0.5B to 7B parameters, using CPT data based on both Fictitious and Realistic knowledge [4]. We then evaluated the knowledge extraction performance under various exposure settings using the methodology described in Section 3.2. Our results show that, across different model sizes and knowledge types, the effectiveness of knowledge extraction consistently follows a log-sigmoid trend with respect to exposure time. In Sections 4.1 and 4.2, we provide a formal definition of this scaling law and identify three distinct phases in its progression. Further analysis in Sections 4.3 and 4.4 explores how this scaling behavior correlates with model scale and dataset characteristics.

4.1 Knowledge Extraction Performance: A Log-Sigmoidal Scaling with Exposure Times

As shown in Figure 1, our experiments reveal that the model's proficiency of knowledge exhibits a log-sigmoid relationship with the number of exposures to the knowledge:

$$P_E(\mathcal{K}; n) = \beta + \frac{\alpha}{(1 + (\frac{n_0}{n})^k)} \tag{3}$$

where n represents the number of exposures to the knowledge in knowledge space $\mathcal{K}$, k controls the steepness of the curve, β denotes the minimum extraction ability for $\mathcal{K}$, α determines the range of the proficiency scaling and n_0 is the inflection point, indicating the exposure times at which the proficiency improves most rapidly. This pattern holds consistently across both fictitious and realistic datasets, suggesting that the observed learning dynamics are generalizable and not tied to any specific data distribution.

4.2 The Three Phases of Knowledge Injection in CPT

As shown in Figure 1, when ordered by the number of exposures from low to high, the sigmoid curve can be roughly divided into three distinct phases: 1) the warmup phase, 2) the rapid learning phase, and 3) the saturation phase.

Formally, given a threshold ratio of the total gain α, denoted by λ [5], the **warmup phase** extends from

[3] Technically, we can select any realistic corpus containing a large amount of factual knowledge. We chose the Middle East Works dataset because its topic strikes a balance between global popularity and regional specificity.

[4] We keep $|\mathcal{K}_F| = |\mathcal{K}_R| = 50,000$ for all experiments and analysis for simplicity, as shown in Figure 6, change the size of knowledge does not affect the log-sigmoid trend.

[5] We set $\lambda = 0.05$ for all results in this paper.

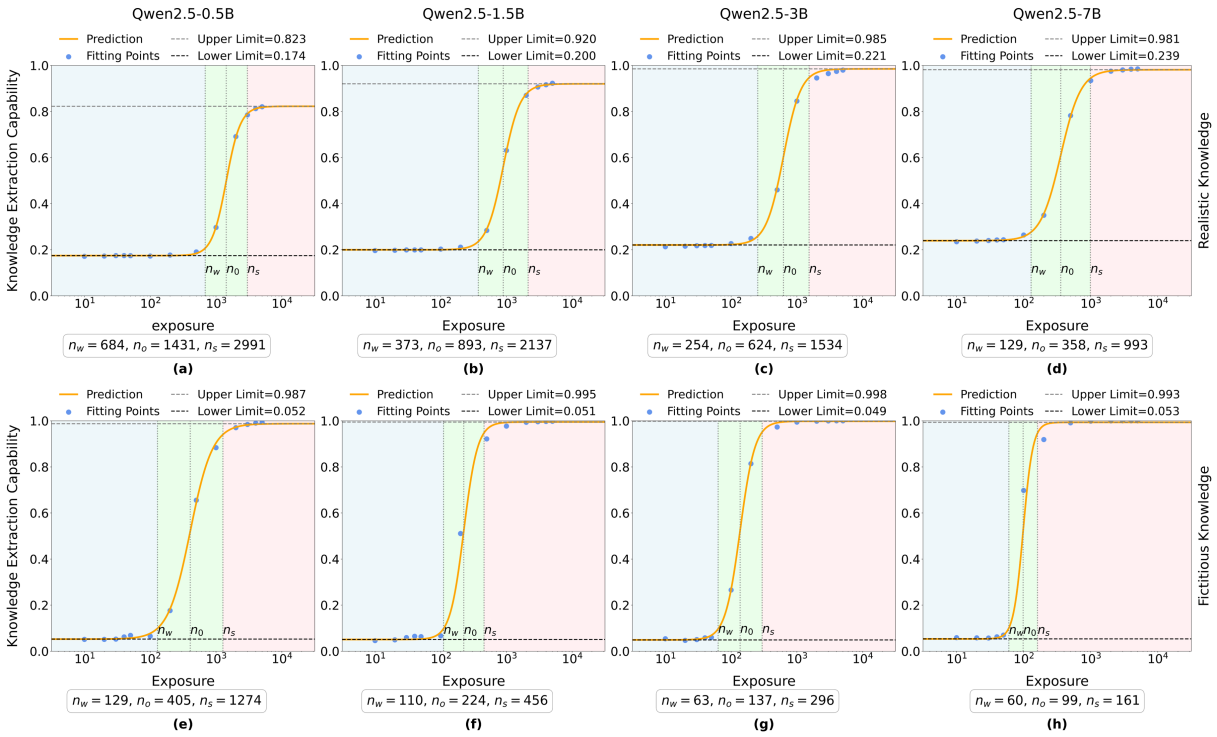


Figure 1: The relationship between LLM knowledge extraction capability and exposure times during CPT: A comparison of models across two datasets. Subfigures (a)-(d) show Middle-East-Works dataset results for 0.5B, 1.5B, 3B, and 7B parameter models, while (e)-(h) display Biography dataset experiments, both dataset contain $50,000$ knowledge. Background shading indicates learning phases: Warmup (blue), Rapid Learning (green), and Saturation (pink). Curves show predicted capability (orange) with actual data points (blue), bounded by asymptotic limits, fitted parameters are presented below the subfigures.

$n = 0$ to the point where performance reaches a fraction λ of the total gain α. That is,

$$P_E(\mathcal{K}; n_w) = \beta + \lambda\alpha \tag{4}$$

Solving for n_w, we obtain:

$$n_w = n_0 \left(\frac{\lambda}{1-\lambda} \right)^{1/k} \tag{5}$$

The **rapid learning phase** refers to the regime in which performance increases sharply with additional exposures. It begins at $n = n_w$ and ends at $n = n_s$, the point at which performance reaches $1 - \lambda$ of the total gain:

$$P_E(\mathcal{K}; n_s) = \beta + (1-\lambda)\alpha \tag{6}$$

Solving for n_s, we get:

$$n_s = n_0 \left(\frac{1-\lambda}{\lambda} \right)^{1/k} \tag{7}$$

Finally, the **saturation phase** begins at $n = n_s$ and continues as $n \to \infty$. During this phase, performance asymptotically approaches its upper bound $\beta + \alpha$, and further improvements become increasingly marginal.

Warmup Phase. During the warmup phase, although the training loss decreases steadily (see Figure 5), the model exhibits little to no improvement in extracting new knowledge, indicating that initial computational effort is spent on domain adaptation rather than actual learning. This behavior mirrors the "undo" effect observed by (Zheng et al., 2025) in early stages of continual fine-tuning, where models first discard old patterns before adapting to new tasks. These findings suggest that the warmup phase serves as a critical realignment process, balancing plasticity and stability to prevent disruptive interference with existing knowledge before meaningful integration can occur.

Rapid Learning Phase. In the rapid learning phase, the model's mastery of the injected knowledge increases most rapidly, exhibiting a log-linear scaling behavior near $n = n_0$, despite only a marginal decrease in training loss compared to the warmup phase. This suggests that computational resources are now primarily allocated to actual knowledge acquisition, rather than domain adap-

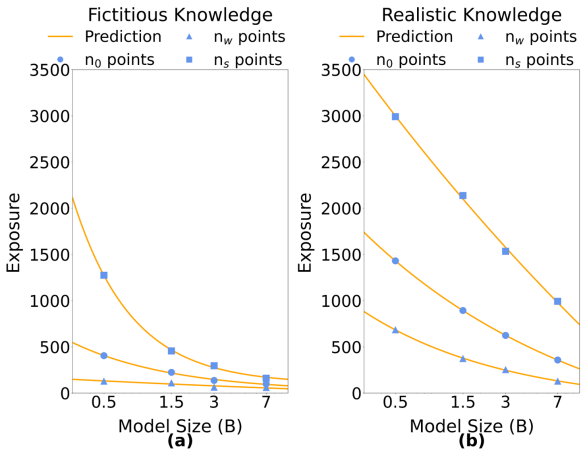


Figure 2: Power law scaling of required exposures $n_\star(M)$ with model size on two knowledge datasets.

tation. Combined with the observations from the warmup phase, this indicates a clear transition in the model's learning dynamics: once the initial realignment of representations is complete, the model enters a regime of efficient knowledge integration, where performance improves rapidly with additional exposure.

Saturation Phase. In the saturation phase, the model's knowledge extraction ability approaches its maximum capacity, reflecting diminishing returns as exposure increases further. This phase highlights the natural limits of knowledge acquisition under the current setup. Notably, for the realistic knowledge set, smaller models such as Qwen2.5-0.5B and Qwen2.5-1.5B reached saturation before approaching near 100% extraction performance, indicating a lower ceiling for knowledge injection in smaller-scale models.

4.3 Exposure Requirements Scale with Model Size via a Power Law

Critical Exposure Requirements. The three critical values n_w, n_0, n_s in the log-sigmoid curve correspond to key exposure requirements during the knowledge injection process of models: specifically, n_w represents the minimum exposure count required for initial adaptation, n_0 marks the exposure level where maximum learning efficiency occurs, and n_s denotes the exposure quantity needed to achieve performance saturation. These metrics enable estimation of training costs for knowledge injection and predictive modeling of achievable performance under fixed computational budgets. Importantly, these exposure requirements are not constant across model sizes. Instead, we observe that larger models typically require fewer exposures to acquire the same knowledge compared to smaller models. This suggests that increased model capacity enhances the efficiency of knowledge absorption, reducing the amount of data or training time needed to reach a given performance level.

Power-Law Scaling Between Exposure Requirements and Model Size. As illustrated in Figure 2, we observe that for specific knowledge sets, these exposure requirements exhibit power-law scaling with model size:

$$n_\star(M) = aM^b + c \tag{8}$$

Where M denotes the number of model parameters, $n_\star$ represent one of n_w, n_0 or n_s, a, b and c are fitted constants. This empirical scaling law reveals a predictable relationship between model size and knowledge exposure demands. The existence of this scaling relationship enables practical applications in resource planning and model development. By measuring exposure requirements on small-scale models, one can extrapolate the expected training costs and performance limits for much larger models. This allows for more informed decision-making in computational investment, supporting efficient prototyping, budget allocation, and predictive modeling of training dynamics.

4.4 Acquiring Realistic Knowledge Is More Challenging Than Fictitious Knowledge

Higher Complexity in Realistic Knowledge. Although one might expect synthetic, unseen knowledge to be harder for models to learn, Figures 1 and Figure 2 clearly demonstrate that realistic knowledge requires significantly more exposure to acquire compared to fictitious knowledge even when both contain the same amount of factual knowledge. This discrepancy suggests a deeper distinction between the two types of knowledge beyond mere authenticity: *diversity*. The fictitious dataset contains only six unique relations, all of which share a common head entity ("name"). In contrast, the realistic dataset includes 19 distinct relations and does not impose such structural uniformity. These differences in relation diversity and structure likely contribute to the increased difficulty in learning realistic knowledge.

Knowledge Diversity Has a Greater Impact Than Familiarity in CPT Knowledge Injection. Figure 3 examines how model parameters vary across different relations within each dataset for the

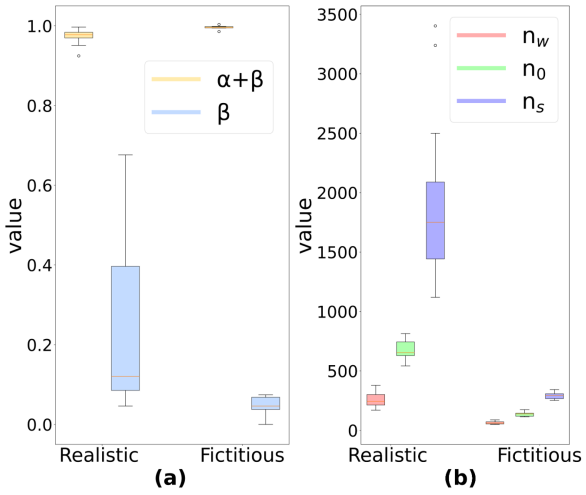


Figure 3: (a) Distribution of baseline (β) and saturated performance levels ($\beta + \alpha$) across different relations under Realistic and Fictitious Knowledge. (b) Distribution of $n_\star$ across different relations under Realistic and Fictitious Knowledge.

Qwen2.5-3B model. As shown in Figure 3 (a), the baseline performance exhibits significantly higher variance on the realistic dataset than on the fictitious one. Since β reflects the model's knowledge extraction capability before injection, this suggests that certain relations in the realistic dataset are inherently more familiar to the pre-trained model, leading to varied performance. This is expected, as modern LLMs are typically pretrained on corpora such as Wikipedia, which contain real-world factual knowledge. In contrast, the fictitious dataset shows little variation, as all knowledge is novel. However, this familiarity does not translate into faster learning, contrary to intuition. As seen in Figure 3 (b), learning realistic knowledge requires significantly more exposures than learning unseen, fictitious knowledge. This suggests that the diversity of knowledge has a greater impact on CPT knowledge injection than its familiarity with the pre-trained model.

Saturation Exposure Varies by Relation Type, but Warmup Exposure Is Robust. As illustrated in Figure 3 (b), the variance of $n_\star$ (including n_w, n_0, n_s) follows a similar trend across relation types. The high variance in the realistic dataset indicates that the amount of exposure required for effective learning varies significantly depending on the specific relation being acquired. In particular, the substantial differences in n_s, the saturation point, suggest that the relation type strongly influences how quickly the model can fully internalize new knowledge. In contrast, the relatively small variance in n_w, which corresponds to the exposures needed during the warmup phase, implies that initial adaptation is less affected by the specific characteristics of each relation. This observation, together with the trend shown in Figure 1, where n_w remains nearly constant across model sizes for fictitious knowledge but decreases notably for realistic knowledge as model size increases, suggests that n_w is largely determined by the model's general real-world knowledge capacity—which improves with scale and is more closely tied to the realistic dataset. Together, these findings indicate that while saturation exposure is highly dependent on relation type, the early phase of adaptation remains relatively consistent across different kinds of knowledge. See Figure 7 and Figure 8 for relation-wise results.

5 Realistic Domain Knowledge Extraction and Data Synthesis

Conducting scaling law research on CPT knowledge injection requires obtaining realistic training data with precise control over both the quantity of knowledge and its exposure times. To tackle this challenge, as illustrated in Figure 4, we developed a framework for data synthesis based on domain-specific corpora. This framework consists of two main steps: a) extraction of factual knowledge from the corpus, and b) synthesis of training data based on the extracted factual knowledge. Section 5.1 details the multi-stage pipeline for extracting high-quality factual knowledge triples from raw corpora by LLMs. Section 5.2 describes the method for synthesizing training data with precisely controlled exposure times using these knowledge triples.

5.1 High-Quality Factual Knowledge Extraction

Defining High-Quality Knowledge Triplets. To support scaling law training and evaluation, we define high-quality factual knowledge triplets based on three criteria: (1) the tail entity must be uniquely inferable from the head and relation; (2) both entities and relations must be clearly and precisely expressed; and (3) the triplet should carry domain-relevant information. We observe that LLMs often extract low-quality triples from open-domain corpora lacking predefined relation scopes—such as ("Mike", "travels to", "New York") or ("Arabic Sands", "is a", "book").

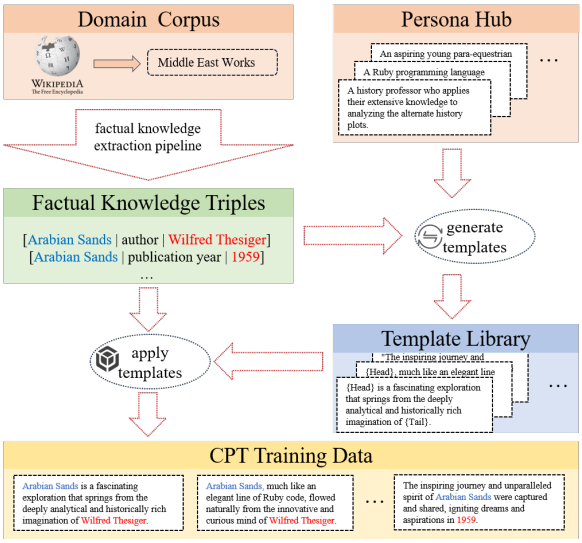


Figure 4: The framework of CPT data synthesis pipeline.

A Multi-Stage Extraction Pipeline. To address this, we designed a four-stage prompting pipeline (see Figure 9) for extracting and refining high-quality triples from Wikipedia. The process begins with Prompt A for initial extraction, followed by Prompt B to remove invalid or implausible triples. Then, Prompt C classifies and standardizes relations into a unified schema, resolving linguistic variations (e.g., "author" vs. "was written by"). Finally, Prompt D re-extracts triples using the refined relation set $\mathcal{R} = \{r\}$ to improve the quality of triples. This multi-stage approach yields a clean, consistent dataset for downstream tasks. Full prompts and implementation details are provided in Appendix D.

5.2 Knowledge based Training Data Synthesis

Given structured knowledge triples $\mathcal{K}_R = \{\mathcal{T}_R\}$, our goal is to synthesize them into natural language training data for CPT, ensuring each fact is exposed multiple times in diverse expressions. Prior work highlights the importance of expression diversity for effective training. However, generating large-scale, semantically coherent variations remains challenging. To address this, we adopt the approach from (Ge et al., 2024), leveraging the extensive persona descriptions in Persona Hub to generate sentence templates tailored to each relation type. This approach enhances linguistic diversity while preserving semantic consistency.

Relation Specified Template Libraries. For each relation $r \in \mathcal{R}$, we construct a prompt using persona descriptions. These prompts are then processed by Qwen2.5-72B-Instruct to generate N [6] unique natural language templates per relation, forming the template library $\mathbf{L}_r$. For instance, for the relation "birth year", example templates include "name was born in year" and "name first appeared in the world in year". These templates enable diverse yet semantically meaningful expressions of factual knowledge. Subsequently, for each triple $\mathcal{T}_R = (h, r, t) \in \mathcal{K}_R$, we apply all corresponding templates in $\mathbf{L}_r$ to generate the final sentences.

6 Conclusion

This study systematically establishes scaling laws of exposures for domain knowledge injection in CPT, identifying two core phenomena: (1) knowledge injection performance follows a log-sigmoid trajectory, and (2) the required exposure scales as a power law with model capacity. These insights provide practical guidance for predicting data synthesis and resource needs in domain-specific training, enabling more efficient use of computational resources. Our new data synthesis framework further offers a flexible and robust tool for studying knowledge injection in real-world settings. This work thus provides both theoretical and practical foundations for next-generation domain-specific language models.

Limitation

Our study investigates the scaling behavior of factual knowledge injection corresponding to exposures during CPT and introduces a data synthesis pipeline; however, several limitations remain: 1) Although we have made progress in creating more realistic synthetic data, a gap still exists between natural corpora and synthesized corpora, and minimizing this gap presents an interesting and meaningful avenue for future research; 2) Due to constraints on computational resources and the availability of pretrained models, our experiments were limited to the Qwen2.5 series, and a broader exploration of scaling laws across different model families is warranted in future work; 3) As this work focuses solely on the efficacy of knowledge injection, the issue of catastrophic forgetting in CPT remains unexplored and should be addressed in future studies.

[6] To avoid confusion with the concept of epochs, we set N to the maximum number of exposure times used in our model training across experiments.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Zeyuan Allen-Zhu and Yuanzhi Li. 2023. Physics of language models: Part 3.1, knowledge storage and extraction. *ArXiv*, abs/2309.14316.

Zeyuan Allen-Zhu and Yuanzhi Li. 2024a. Physics of language models: Part 3.3, knowledge capacity scaling laws. *ArXiv*, abs/2404.05405.

Zeyuan Allen-Zhu and Yuanzhi Li. 2024b. Physics of language models: Part 3.3, knowledge capacity scaling laws. *arXiv preprint arXiv:2404.05405*.

M Saiful Bari, Yazeed Alnumay, Norah A Alzahrani, Nouf M Alotaibi, Hisham Abdullah Alyahya, Sultan AlRashed, Faisal Abdulrahman Mirza, Shaykhah Z Alsubaie, Hassan A Alahmed, Ghadah Alabduljabbar, et al. 2025. Allam: Large language models for arabic and english. In *The Thirteenth International Conference on Learning Representations*.

Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, and Minjoon Seo. 2024. How do large language models acquire factual knowledge during pretraining? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Bap tiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Cantón Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab A. AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriele Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guanglong Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Laurens Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Ju-Qing Jia, Kalyan Vasuden Alwala, K. Upasani, Kate Plawiak, Keqian Li, Ken-591 neth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuen ley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Babu Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melissa Hall Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri S. Chatterji, Olivier Duchenne, Onur cCelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasić, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Ro main Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Chandra Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whit ney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yiqian Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zhengxu Yan, Zhengxing Chen, Zoe Papakipos, Aaditya K. Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adi Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Ben Leonhardi, Po-Yao (Bernie) Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Shang-Wen Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Mont-

gomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzm'an, Frank J. Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory G. Sizov, Guangyi Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Han Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kaixing(Kai) Wu, U KamHou, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, A Lavender, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollár, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sung-Bae Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Andrei Poenaru, Vlad T. Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xia Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Jiawei Gu, Zacc Yang, Chuanghao Ding, Rui Zhao, and Fei Tan. 2024. Cmr scaling law: Predicting critical mixture ratios for continual pre-training of language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16143–16162.

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*. Association for Computational Linguistics.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, et al. 2024. Acegpt, localizing large language models in arabic. In *NAACL-HLT*.

Juhao Liang, Zhenyang Cai, Jianqing Zhu, Huang Huang, Kewei Zong, Bang An, Mosen Alharthi, Juncai He, Lian Zhang, Haizhou Li, et al. 2024. Alignment at pre-training! towards native alignment for arabic llms. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun-Qing Li, Hejie Cui, Xuchao Zhang, Tian yu Zhao, Amit Panalkar, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, Chris White, Quanquan Gu, Jian Pei, Carl Yang, and Liang Zhao. 2023. Domain specialization as the key to make large language models disruptive: A comprehensive survey.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. 2024b. Best practices and lessons learned on synthetic data for language models. *CoRR*.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11065–11082.

Xingyu Lu, Xiaonan Li, Qinyuan Cheng, Kai Ding, Xuan-Jing Huang, and Xipeng Qiu. 2024. Scaling laws for fact memorization of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11263–11282.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. 2023. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376.

Haoran Que, Jiaheng Liu, Ge Zhang, Chenchen Zhang, Xingwei Qu, Yinghao Ma, Feiyu Duan, Yuanxing Zhang, Xu Tan, Jie Fu, et al. 2024. D-cpt law: Domain-specific continual pre-training scaling law for large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, and Hao Wang. 2024. Continual learning of large language models: A comprehensive survey. *ArXiv*, abs/2404.16789.

Zirui Song, Bin Yan, Yuhan Liu, Miao Fang, Mingzhe Li, Rui Yan, and Xiuying Chen. 2025. Injecting domain-specific knowledge into large language models: A comprehensive survey. *ArXiv*, abs/2502.10708.

Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Nemotron-cc: Transforming common crawl into a refined long-horizon pretraining dataset. *arXiv preprint arXiv:2412.02595*.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Towards building open-source language models for medicine.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candes, and Tatsunori Hashimoto. 2025b. Synthetic continued pretraining. In *The Thirteenth International Conference on Learning Representations*.

Junhao Zheng, Xidi Cai, Shengjie Qiu, and Qianli Ma. 2025. Spurious forgetting in continual learning of language models. In *The Thirteenth International Conference on Learning Representations*.

Wenzhen Zheng, Wenbo Pan, Xu Xu, Libo Qin, Li Yue, and Ming Zhou. 2024. Breaking language barriers: Cross-lingual continual pre-training at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7725–7738.

A Training loss of CPT

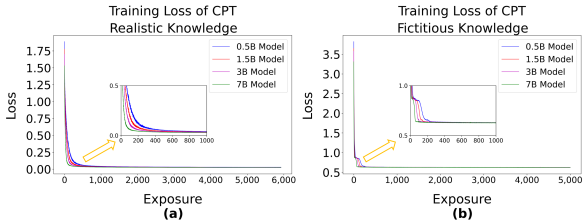


Figure 5: Training loss of CPT for models of 0.5B (blue), 1.5B (red), 3B (purple), and 7B (green) parameters

B Training Details

CPT Training Setup. We set the learning rate to $7e-6$ for all experiments. For data with different exposure times, we used different global batch size values to ensure sufficient updates during the training process, specifically, for $n = 10$, 50, and 100, we conducted separate training sessions with a global batch of 32 instead of 96 used for larger exposures. In our experiment, the average number of tokens per data sample is 32, with the maximum sequence length set to $2,048$. When performing data concatenation, we followed the approach used in DeepSeek-V3 (Liu et al., 2024a) to ensure the integrity of the content was preserved. More hyperparameters are shown in Table 2.

Supervised Fine-Tuning Setup. For the Fine-tuning process described in Section 3.2, we employed a learning rate corresponding to 10% of the original learning rate used in the CPT process, maintaining a global batch size of 96 across all experiments. Through systematic experimentation

Table 2: The list of hyperparameters.

Hyperparameters	Value
Warm-up Steps	0
Gradient Accumulation Steps	2
Max Sequence Length	2048
Learning Rate	7e-6
Min Learning Rate	7e-7
Learning Rate Scheduler	cosine with min lr

with varying epoch numbers, our results demonstrated that the model achieved optimal QA performance at 2 training epochs. This configuration was therefore selected as the optimal training duration, yielding peak performance metrics in our evaluations.

C More Exposure Scaling Results of Knowledge Injection in CPT

Here we present the exposure scaling results comparing different number facts in Figure 6, and different relation types in Figure 7 and Figure 8.

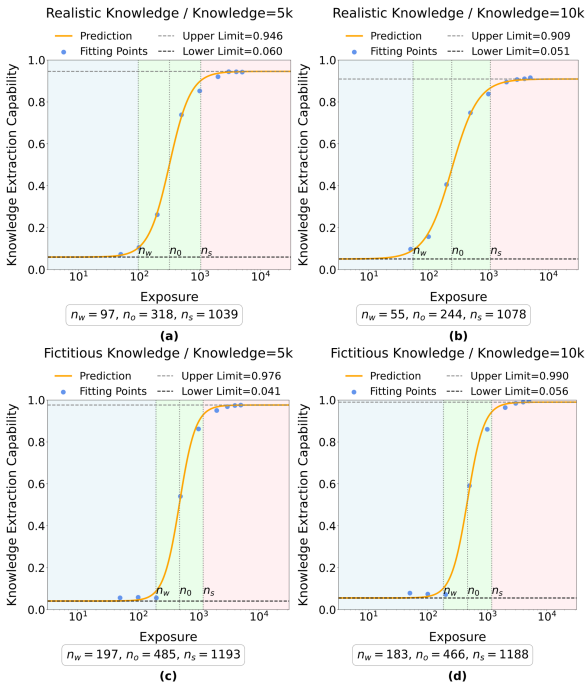
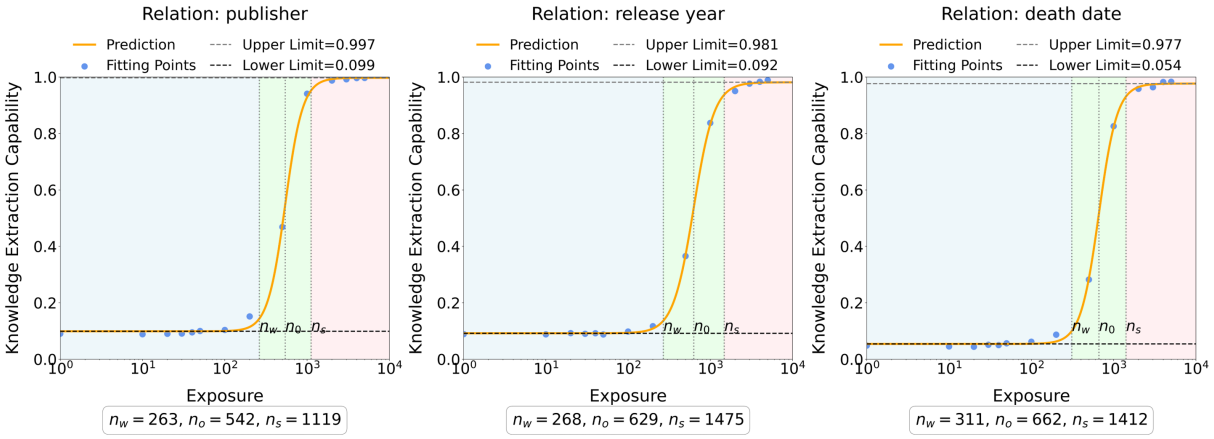


Figure 6: The relationship between LLM knowledge extraction capability and exposure times during CPT on Qwen2.5-0.5B across knowledge sizes of $10,000$ and $5,000$.

D Details of Factual knowledge extraction pipeline.

The factual knowledge extraction pipeline is shown in Figure 9, and all the related prompts are shown in Figure 10 to Figure 14.



$n_w = 263$, $n_o = 542$, $n_s = 1119$

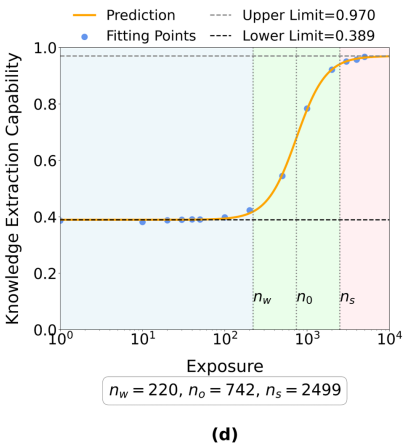
(a)

$n_w = 268$, $n_o = 629$, $n_s = 1475$

(b)

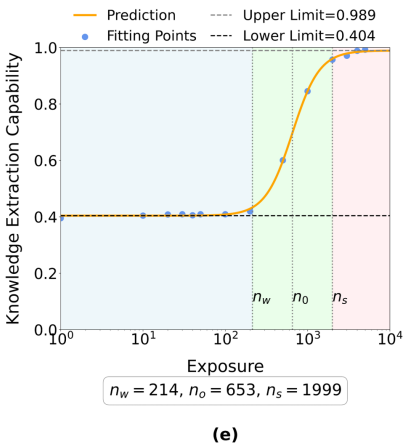
$n_w = 311$, $n_o = 662$, $n_s = 1412$

(c)



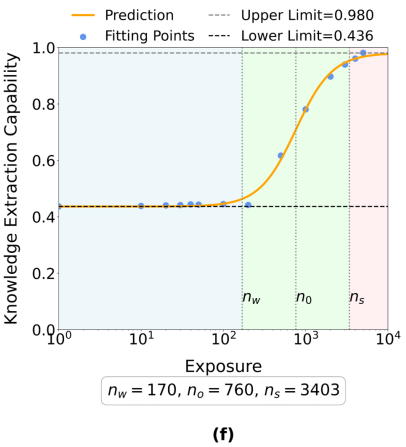
$n_w = 220$, $n_o = 742$, $n_s = 2499$

(d)



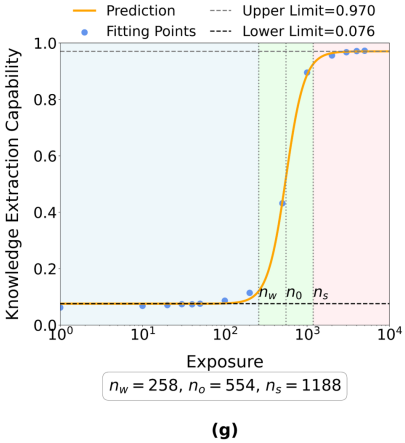
$n_w = 214$, $n_o = 653$, $n_s = 1999$

(e)



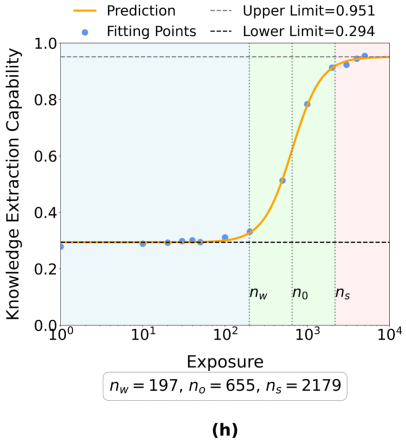
$n_w = 170$, $n_o = 760$, $n_s = 3403$

(f)



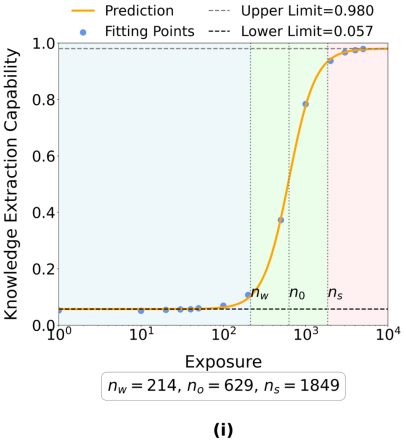
$n_w = 258$, $n_o = 554$, $n_s = 1188$

(g)



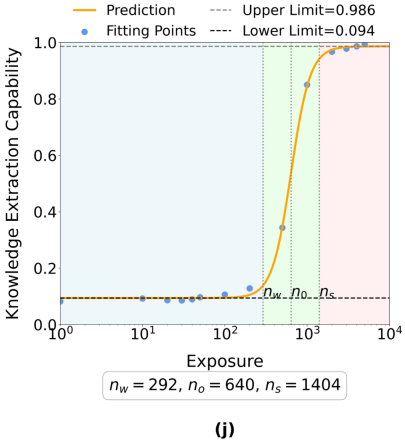
$n_w = 197$, $n_o = 655$, $n_s = 2179$

(h)



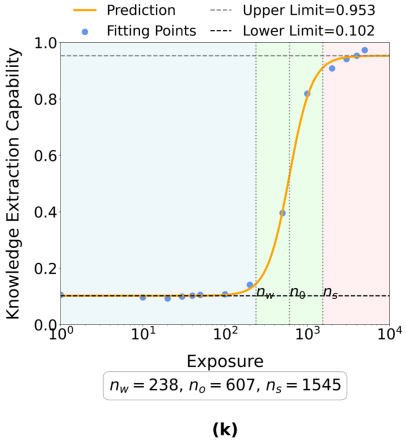
$n_w = 214$, $n_o = 629$, $n_s = 1849$

(i)



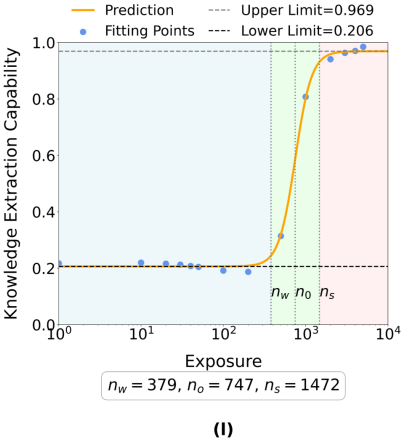
$n_w = 292$, $n_o = 640$, $n_s = 1404$

(j)



$n_w = 238$, $n_o = 607$, $n_s = 1545$

(k)



$n_w = 379$, $n_o = 747$, $n_s = 1472$

(l)

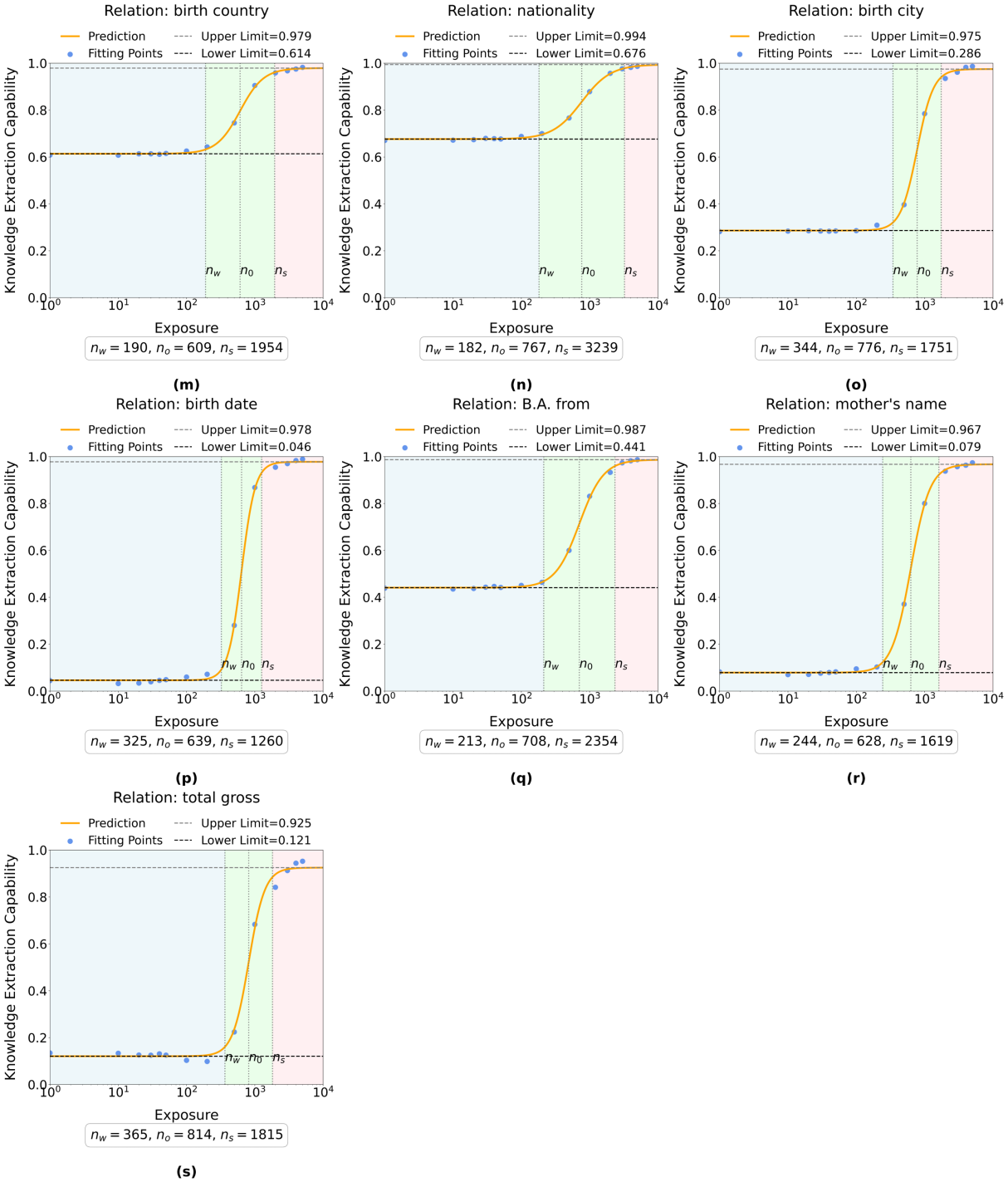


Figure 7: The relationship between LLM knowledge extraction capability and exposure times during CPT on Qwen2.5-3B for different relations in realistic dataset.

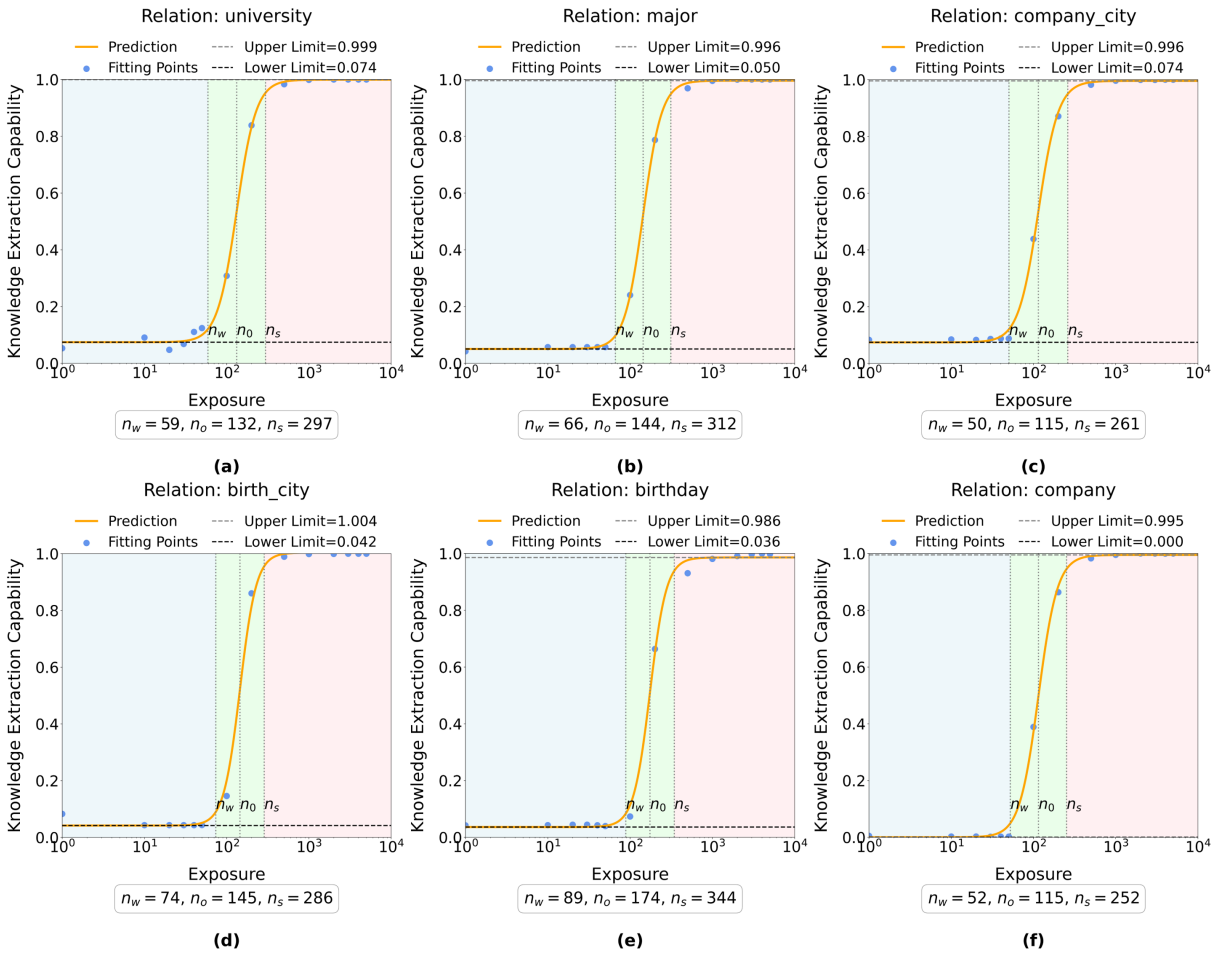


Figure 8: The relationship between LLM knowledge extraction capability and exposure times during CPT on Qwen2.5-3B for different relations in fictitious dataset.

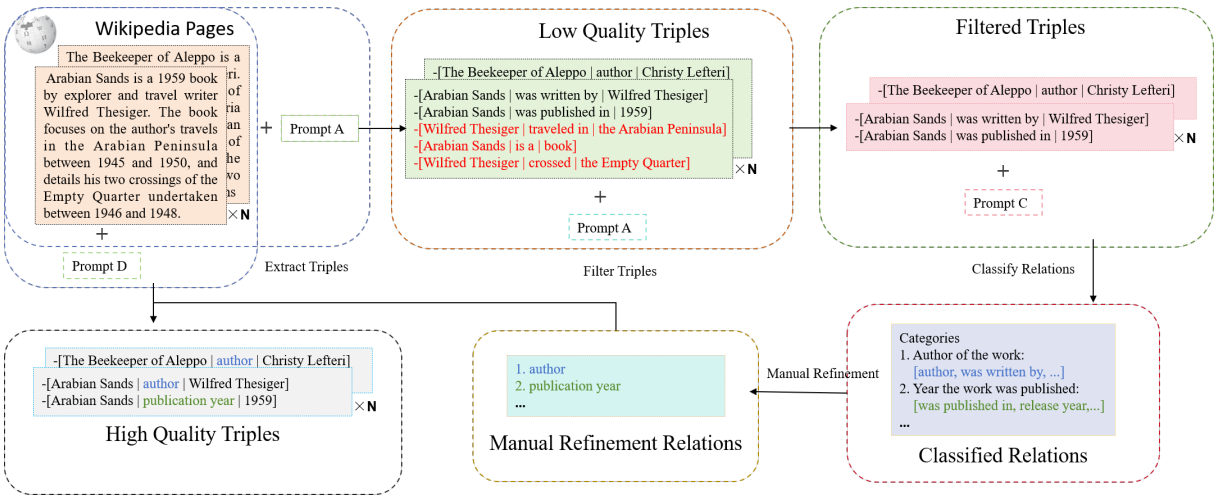


Figure 9: Factual knowledge extraction pipeline. The process begins by extracting low-quality triples from Wikipedia pages using Prompt A. These triples are filtered using Prompt B to remove invalid triples (red-highlighted examples). The filtered triples are then categorized based on their relations using Prompt C, such as "author" (blue) and "publication year" (green). Manual refinement unifies variations of the same relation within each category. These refined relations are embedded into Prompt D to re-extract high-quality, standardized triples from the original pages, ensuring structured and accurate factual knowledge construction.

Prompt A: Low-Quality Triples Extraction

Extract factual knowledge triples from the text below. Follow these rules:

1. Only include static facts (e.g., dates, authorship, locations).
2. Format each triple as: -[Head Entity | Relationship | Tail Entity], which is equivalent to [Subject | Predicate | Object].
3. Extract at least 20 triples.
4. No explanations needed.

Text:
```
{text}
```

Output format:
```
-[Entity 1 | relationship 1 | Entity 2]
-[Entity 3 | relationship 2 | Entity 4]
```

Figure 10: Prompt for extracting low-quality triples from Wikipedia pages.

Prompt B: Triples Filtering Prompt

Analyze whether each extracted triple represents a **unique factual relationship** where the tail entity has no other possible values for the given head entity and relationship. Follow these steps:

1. For each triple, check:

- If the tail entity **must be unique** (e.g., publication year, locations).
- Exclude ambiguous relationships (e.g., "crossed by", professions, "travels to", "moved to").
- Fix incorrect triples by swapping head/tail entities if logically inverted.

2. Examples:

Invalid

1. [Wilfred Thesiger | profession | explorer] -> Invalid. "profession" allows multiple values.
2. [Aziz Nesin | created character | Zübük] -> Invalid. Head/tail inversion because "Zübük" is not the only valid value for the tail entity when head is "Aziz Nesin" and the relation is "created character".
   - Correction: [Zübük | created by | Aziz Nesin], "Aziz Nesin" is the only valid value for the tail entity in this triple.
3. [The Image Book | Award | Special Palme d'Or] -> Invalid. The Image Book has won more than one award. "Special Palme d'Or" could be replaced by others.
4. [Brush teeth | timeframe | 8:00 AM] -> Invalid. The entity "Brush teeth" is ambiguous without specifying who performed the brushing.
5. [J.K. Rowling | wrote | Harry Potter] -> Invalid. Head/tail inversion because "Harry Potter" is not the only valid value for the tail entity when the head is "J.K. Rowling" and the relation is "wrote".
   - Correction: [Harry Potter | written by | J.K. Rowling], "J.K. Rowling" is the only valid value for the tail entity in this triple.
6. [Mike | travels to | New York] -> Invalid. Mike may travels to other cities, not only "New York" can be the valid value for the tail entity.

Valid

1. [Manwakh | located in | Yemen] -> Valid. "located in" is fixed.
2. [TCP/IP | publication year | 1974] -> Valid. "Publication years" are singular factual events.

3. Analyze each triple below:

Triples to validate:
{triples}

Output format:

Analysis:

1.[Triple 1] → [Valid/Invalid]. *[Brief reason]*.
 - Correction: `[New Head | New Relation | New Tail]` (if applicable)
2.[...]

The Valid/Corrected Triples:
```
-[Head | Relation | Tail]
-[...]
```

Figure 11: Triples Filtering Prompt: Steps and examples for analyzing and verifying unique factual relations. In this process, each triple is examined to determine whether the tail entity is unique for a given head entity and relation, meaning that the tail entity cannot have alternative possible values.

Prompt C: Triples Classification Prompt

You are a knowledge graph expert. I will provide you with some triples below. These triples involve many categories. Please help me summarize how these triples can be categorized based on their relations. For each category, please output a few of the triples I provided as examples. Place the categories with a higher proportion at the top.

Triples:
{triples}

Figure 12: Triples Classification Prompt: Summarize relation classes and provide examples.

Prompt D: High-Quality Triples Extraction

Please extract triples in the form of (Entity1, Relation, Entity2) from the text provided below. Ensure that each "Relation" is strictly selected from the predefined list of relations provided. If no matching relation can be found in the text based on the predefined list, output `None`.

Predefined Relations:
- **author**: Indicates that Entity2 is the author of Entity1.
  - *Example*: `["Harry Potter", "author", "J.K. Rowling"]` means J.K. Rowling is the author of Harry Potter.
- **director**: Indicates that Entity2 is the director of Entity1.
  - *Example*: `["A", "director", "B"]` means B is the director of A.
- **creater**: Indicates that Entity2 is the creater of Entity1.
- **birth date**: Represents the birth date of Entity1.
  - *Example*: `["Mike", "birth date", "January 1, 1990"]`
- **birth year**: Represents the birth year of Entity1.
  - *Example*: `["Mike", "birth year", "1990"]`
...

Triple Extraction Examples
Text:
```
_**Willow and Wind**_ (Persian: _Beed-o baad_) is a 2000 Iranian drama film directed by Mohammad-Ali Talebi and written by Abbas Kiarostami.

## Cast

* Dariush Afshar as Soraya Esfandiari  *Arman Naderi as Yasmin Khorrami
```

Output:
```json
{
  "triples": [
    ["Willow and Wind", "director", "Mohammad-Ali Talebi"],
    ["Willow and Wind", "author", "Abbas Kiarostami"],
    ["Willow and Wind", "made in", "Iran"],
    ["Willow and Wind", "release year", "2000"],
  ]
}
```

Text for Analysis:
```{content} ```

Please return the results in JSON format as follows:
```json
{
  "triples": [
    [Entity1, relation, Entity2],
    [...]
  ]
}
```

If no triples can be extracted based on the predefined relations, please output:
```json
{
  "triples": null
}
```

Figure 13: High-quality Triples Extraction and Classification: Extracting triples from text based on a predefined list of 26 relation types (partially shown in the figure for brevity). Relations include: B.A. from, Ph.D. from, academic advisor, author, birth city, birth country, birth date, birth year, creator, death date, death year, director, father's name, located in, made in, master's degree from, mother's name, nationality, portrayed by, publish year, publisher, release by, release date, release year, total gross, wife's name. Each extracted triple strictly adheres to this predefined schema.
```

Prompt E: Template Generation Prompt

Assume you are the persona described below, and you are crafting a sentence in the persona's style to describe the relationship between a person and the specific date of their birth.

Requirements:

1.placeholders such as {Head} and {Tail} should be used.
2.The output should be in English.

Persona:
an IT project manager who adopted extreme programming (XP) methodologies on his own team.
Output:
{Head} came into existence on the timeline of life on {Tail}, marking the starting point of their journey.

Persona:
A nature photographer who wants to showcase their stunning photographs with sustainable and unique frames
Output:
{Head} entered the world on the beautiful day of {Tail}, a moment that would inspire a lifetime of capturing nature's splendor.

Persona:
{persona}
Output:

Figure 14: Template Generation Prompt: Generate sentences in the style of a specific person that can be filled with head and tail entities (using the relation between a person and their birth date as an example). To ensure diversity in the generated templates, allow the use of statements similar to the relation in the template for substitution.

Template Examples

Persona: A paralyzed individual who hopes to regain some motor control through brain-computer interface therapy.
Template: {Head} was born on the significant date of {Tail}, initiating a life marked by resilience and the pursuit of groundbreaking advancements in brain-computer interface therapy.

Persona: A blogger who writes in-depth reviews and reflections on each monthly read.
Template: {Head} embarked on their life's narrative on the page of time known as {Tail}, setting the stage for a lifetime of turning the pages of countless stories.

Figure 15: Template Example: Sentences describing entity relations in the style of a specific person (using the relation between a person and their birth date as an example). Fill in the person's name at Head and the birth date at Tail.

Synthesis Data Examples

Data: Saul Bellow was born on the significant date of April 5, 2005, initiating a life marked by resilience and the pursuit of groundbreaking advancements in brain-computer interface therapy.
Head: Saul Bellow
Relation: death date
Tail: April 5, 2005

Data: Bernard Lewis embarked on their life's narrative on the page of time known as May 19 2018, setting the stage for a lifetime of turning the pages of countless stories.
Head: Bernard Lewis
Relation: death date
Tail: May 19, 2018

Figure 16: Synthetic Data Example. (using the relation between a person and their birth date as an example)