Prior Knowledge Shapes Fine-Tuning Success for Biomedical Term Normalization

Daniel B Hier *, Steven Keith Platt, Anh Nguyen

Posted Date: 7 August 2025

doi: 10.20944/preprints202508.0574.v1

Keywords: large language model; fine-tuning; LoRA; PEFT; human phenotype ontology; biomedical terminology; knowledge injection; natural language processing; term normalization; ontology li

 doi:10.20944/preprints202508.0574.v1

Article

Prior Knowledge Shapes Fine-Tuning Success for Biomedical Term Normalization

Daniel B. Hier [1,*], Steven Keith Platt [2], Anh Nguyen [2]

[1] Dept. of Neurology & Rehabilitation, University of Illinois at Chicago, Chicago, IL USA
[2] Laboratory for Applied Artificial Intelligence, Loyola University Chicago, Chicago, IL USA
* Correspondence: dhier@uic.edu

Abstract

Large language models (LLMs) often fail to correctly link biomedical terms to their standardized ontology identifiers, posing challenges for downstream applications that depend on accurate, machine-readable codes. These linking failures can compromise the integrity of data used in precision medicine, clinical decision support, and population health. Fine-tuning can partially remedy these issues, but the degree of improvement varies across terms and terminologies. Focusing on the Human Phenotype Ontology (HPO), we show that a model's prior knowledge of term–identifier pairs, acquired during pre-training, strongly predicts whether fine-tuning will enhance its linking accuracy. We evaluate prior knowledge in three complementary ways: (1) Latent probabilistic knowledge, revealed through stochastic prompting, captures hidden associations not evident in deterministic output; (2) Partial subtoken knowledge, reflected in incomplete but non-random generation of identifier components; and (3) Term familiarity, inferred from annotation frequencies in the biomedical literature, which serve as a proxy for training exposure. We then assess how these forms of prior knowledge influence deterministic accuracy in identifier linking. Fine-tuning performance varies most for terms in what we call the reactive middle zone of the ontology—terms with intermediate levels of prior knowledge that are neither absent nor fully consolidated. These terms exhibit the largest gains or losses in accuracy during fine-tuning, suggesting that the success of knowledge injection critically depends on the initial level of model familiarity with the term–identifier pair.

Keywords: large language model; fine-tuning; LoRA; PEFT; human phenotype ontology; biomedical terminology; knowledge injection; natural language processing; term normalization; ontology linking

1. Introduction

Biomedical term normalization—the process of mapping natural language expressions to standardized ontology concepts with machine-readable identifiers—is a cornerstone of precision medicine and biomedical research. Accurate normalization enables clinical and scientific text to be aligned with structured knowledge resources, supporting reproducible analyses and computational reasoning.

Widely adopted ontologies include SNOMED CT, which provides standardized terminology for patient care, and the Gene Ontology (GO), which supports functional annotation in gene and protein research. The Human Phenotype Ontology (HPO), a structured vocabulary for describing phenotypic abnormalities associated with human disease [1–3], plays a particularly important role in elucidating genetic disease mechanisms. With more than 18,000 terms and over 268,000 disease annotations, each linking a phenotypic feature and its ontology identifier to a specific disease, the HPO connects phenotypic, genomic, and disease knowledge. This integration enables large-scale biomedical research into the genetic basis of human disease [4].

Successful use of the HPO for disease annotation relies on the accurate linking of each phenotype term to its corresponding ontology identifier [2]. For example, Charcot–Marie–Tooth disease type 2B2 is annotated with phenotypic features such as *Distal muscle weakness* (HP:0002460), *Areflexia* (HP:0001284),

 doi:10.20944/preprints202508.0574.v1

and *Distal sensory impairment* (HP:0002936). Each HPO term is expressed in title case and uniquely identified by a seven-digit code prefixed with HP:. Precise alignment of the term–identifier pair is essential for ensuring semantic accuracy and interoperability, enabling downstream applications such as diagnostic decision support, cohort discovery, and automated biomedical data analysis.

Large language models (LLMs) are increasingly used in biomedical natural language processing (NLP) tasks such as entity recognition, relation extraction, and term normalization [5]. Despite their broad success, LLMs often struggle with the seemingly straightforward task of biomedical term normalization [6]. Even state-of-the-art models frequently fail to link Human Phenotype Ontology (HPO) terms to their correct identifiers. A core limitation lies in the pre-training paradigm: autoregressive LLMs such as GPT-4 and LLaMA 3.1 are optimized for next-token prediction, not explicit fact memorization. While these models are exposed to vast biomedical corpora during pre-training, many rare biomedical concepts remain sparsely represented. As a result, model performance drops sharply across the long tail of biomedical vocabularies, where domain-specific terms are underexposed [7]. For instance, when GPT-4 was queried for identifiers corresponding to 18,880 HPO terms, it returned the correct identifier for only 8% [6].

Fine-tuning offers a potential remedy. Targeted adaptation using parameter-efficient methods such as LoRA has shown promise for injecting missing knowledge into large language models (LLMs) [8–12]. However, recent evaluations caution that these gains are often inconsistent: smaller models struggle to generalize reliably, and improvements in newly injected content may come at the cost of degrading existing knowledge—particularly when that knowledge is fragile or only weakly anchored [8,13–18].

A growing body of work suggests that while fine-tuning enables models to memorize new term–identifier mappings, it instills a limited ability to generalize from these mappings to unseen expressions [16,17]. Rather than promoting conceptual integration, fine-tuning may act as a form of rote injection, reinforcing isolated facts without building robust representations. Consequently, the success of fine-tuning appears to depend not only on the added data but also on how well the target concept is already embedded in the model's pretraining knowledge [17,19]. The jury is still out on whether certain biomedical ontologies, by virtue of their innate structure or prevalence of their concepts in training data, may support broader conceptual generalization during fine-tuning than other less robust ontologies.

This raises a core question: Which biomedical terms are most likely to benefit from fine-tuning, and which are most vulnerable to degradation? We hypothesize that both improvement and degradation are systematically shaped by the model's prior knowledge of each term–identifier pair.

To test this hypothesis, we evaluate the predictive value of three dimensions of prior knowledge, defined as the model's ability to produce or approximate correct identifier mappings before fine-tuning:

1. *Latent probabilistic knowledge*: Hidden or partially accessible knowledge revealed through probabilistic querying of the model, even when deterministic (greedy) decoding fails [20]. For example, if an LLM is queried 100 times and returns the correct ontology ID only 5% of the time, this indicates Latent probabilistic knowledge, distinct from not known at all.
2. *Partial subtoken knowledge*: Incomplete but non-random knowledge of the subtoken sequences comprising ontology identifiers, reflected in deterministic outputs that are close to, but not exactly, correct. For example, an LLM that predicts HP:0001259 for Ataxia instead of the correct HP:0001251 demonstrates Partial subtoken knowledge, even though greedy decoding produces an incorrect response.
3. *Term familiarity*: The likely exposure of the model to specific term–identifier pairs during pretraining, estimated using external proxies such as annotation frequency in OMIM and Orphanet [21,22], and identifier frequency in the PubMed Central (PMC) corpus [23]. For example, the LLM is more likely to be familiar with the term *Hypotonia* (decreased tone), which has 1,783 disease annotations in HPO, than with *Mydriasis* (small pupils), which has only 25 annotations.

 doi:10.20944/preprints202508.0574.v1

Across these three dimensions, Latent probabilistic knowledge, Partial subtoken knowledge, and Term familiarity, we uncover a striking pattern: terms with intermediate levels of prior knowledge, neither fully consolidated nor entirely absent [20], are the most responsive to fine-tuning. These reactive middle terms exhibit both the largest improvements and the greatest degradations, whereas terms at the extremes remain more stable and less affected by fine-tuning. This dual effect suggests that fine-tuning is both most effective and most disruptive in regions of moderate prior knowledge. Our findings extend prior work by Pletenev et al. [17] and Gekhman et al. [19,20], and underscore the importance of considering term susceptibility to fine-tuning when attempting knowledge injection.

This raises a core question: Which biomedical terms are most likely to benefit from fine-tuning, and which are most vulnerable to degradation? We hypothesize that both improvement and degradation are systematically shaped by the model's prior knowledge of each term–identifier pair.

To test this hypothesis, we evaluate the predictive value of three dimensions of prior knowledge, defined as the model's ability to produce or approximate correct identifier mappings before fine-tuning:

1. *Latent probabilistic knowledge*: Hidden or partially accessible knowledge revealed through probabilistic querying of the model, even when deterministic (greedy) decoding fails [20]. For example, if an LLM is queried 100 times and returns the correct ontology ID only 5% of the time, this indicates Latent probabilistic knowledge, distinct from not known at all.
2. *Partial subtoken knowledge*: Incomplete but non-random knowledge of the subtoken sequences comprising ontology identifiers, reflected in deterministic outputs that are close to, but not exactly, correct. For example, an LLM that predicts HP:0001259 for Ataxia instead of the correct HP:0001251 demonstrates Partial subtoken knowledge, even though greedy decoding produces an incorrect response.
3. *Term familiarity*: The likely exposure of the model to specific term–identifier pairs during pretraining, estimated using external proxies such as annotation frequency in OMIM and Orphanet [21,22], and identifier frequency in the PubMed Central (PMC) corpus [23]. For example, the LLM is more likely to be familiar with the term *Hypotonia* (decreased tone), which has 1,783 disease annotations in HPO, than with *Mydriasis* (small pupils), which has only 25 annotations.

Across these three dimensions—latent probabilistic knowledge, partial subtoken knowledge, and term familiarity—we uncover a striking pattern: terms with intermediate levels of prior knowledge, neither fully consolidated nor entirely absent [20], are the most responsive to fine-tuning. These reactive middle terms exhibit both the largest improvements and the greatest degradations, whereas terms at the extremes remain more stable and less affected. This dual effect suggests that fine-tuning is both most effective and most disruptive in regions of moderate prior knowledge. Our findings extend prior work by Pletenev et al. [17] and Gekhman et al. [19,20], and underscore the importance of considering a term's amenability to fine-tuning when designing knowledge injection strategies.

The remainder of this paper is organized as follows: Section 2 describes the materials and methods, including the experimental setup, probabilistic querying protocol, subtoken analysis, and familiarity scoring. Section 3 presents the results of our evaluation of fine-tuning performance and the predictive value of the three dimensions of prior knowledge. Section 4 discusses the implications of these findings, including how latent knowledge and term amenability shape fine-tuning outcomes, and outlines directions for future research. Section 5 concludes with practical recommendations for designing effective knowledge injection strategies for biomedical term normalization.

2. Materials and Methods

2.1. HPO Dataset

We obtained the Human Phenotype Ontology (HPO) in CSV and OBO formats from BioPortal (https://bioportal.bioontology.org/ontologies/HP) and the OBO Foundry (http://purl.obolibrary.org/obo/hp.obo), respectively. As of May 5, 2025, the ontology comprised of 23,065 classes (18,988 were active phenotype terms used for disease annotations). Each entry includes the standardized term label, unique HPO identifier, hierarchical parent–child relationships, synonyms, and definitions.

 doi:10.20944/preprints202508.0574.v1

Phenotype–disease associations were obtained from the HPO annotation file (https://hpo.jax.org/data/annotations), containging 272,061 annotations that span 12,691 diseases. Most annotations are derived from OMIM and Orphanet. On average, each disease has 21 phenotype annotations.

2.2. Test Terms

To evaluate model performance on samples of realistic clinical language, we curated a test set of 799 HPO term–identifier pairs extracted from 1,618 de-identified neurology physician notes from the University of Illinois electronic health record (EHR) system [24]. Use of these notes was approved by the Institutional Review Board (IRB) of the University of Illinois at Chicago. The clinical term extraction process used a natural language processing pipeline described in [24], which employed GPT-4 prompting strategies to identify 2,718 candidate phrases that corresponded to HPO concepts. The prompts were designed to capture both explicit phenotype mentions and implicit clinical descriptions. After filtering and mapping, 799 unique HPO term–identifier pairs forme the test set for model evaluation.

2.3. Large Language Models and Fine-Tuning

We used LLaMA 3.1 8B as our base model, a transformer-based autoregressive language model with 8 billion parameters pretrained on a diverse mixture of web data, books, code, and scientific text. All experiments used consistent software versions and configuration parameters to ensure reproducibility. Fine-tuning was performed locally using Hugging Face Transformers with Unsloth and Low-Rank Adaptation (LoRA) [25], a parameter-efficient approach that inserts rank-decomposed matrices into attention layers without modifying base weights. While LoRA reduces the number of trainable parameters and can preserve prior capabilities, catastrophic forgetting can still occur [26,27]. Training was conducted on a multi-GPU workstation equipped with three NVIDIA Quadro RTX 8000 GPUs (48 GB VRAM each), CUDA 12.0, and mixed precision (fp16) for optimized memory usage. The model was fine-tuned on the full HPO vocabulary (18,988 terms) using five prompt variations per term–identifier pair. Following Unsloth's recommendations, fine-tuning was limited to three epochs to mitigate overfitting.

2.4. Deterministic Accuracy

We measured deterministic accuracy by prompting models to return the HPO identifier for a given term:

```
prompt = "What is the HPO ID for {term}?
          Return only the code in format HP:1234567"
```

A response was scored as **correct** if the predicted identifier exactly matched the ground truth HPO ID; otherwise, it was marked as **incorrect**.

2.5. Latent Probabilistic Knowledge

Latent probabilistic knowledge was quantified following the method of Gekhman et al. [20]. The base model was prompted 50 times for each term at a temperature of 1.0, enabling diverse probabilistic outputs. Probabilistic accuracy (proportion of correct responses) was computed for each term. Model latent probabilistic knowledge of each term was categorized as:

`Unknown`: probabilistic accuracy = 0.0
`Weak`: $0.0 <$ accuracy < 0.2
`Medium`: $0.2 \leq$ accuracy < 0.5
`Strong`: accuracy ≥ 0.5

Figure 1 shows the distribution of probabilistic accuracy (log-scaled y-axis).

 doi:10.20944/preprints202508.0574.v1

2.6. Partial Subtoken Knowledge

Partial subtoken knowledge was the model knowledge of each of the three HPO identifier subtokens during deterministic sampling. The LLaMa 3.1 tokenizer extracts three subtokens from the 7 digits of the HPO identifier (e.g., `HP:1234567`) is tokenized into three numeric subtokens in the format of 123-456-7. Deterministic model outputs were scored from 0–3 based on the number of correctly predicted numeric subtokens. The four categories of partial subtoken knowledge were:

`None`: 0 of 3 correct
`Weak`: 1 of 3 correct
`Medium`: 2 of 3 correct
`Complete`: 3 of 3 correct

Figure 2 shows the counts of terms in each of the partial subtoken knowledge categories.

2.7. Term Familiarity

We hypothesize that a model's familiarity with term–identifier pairs is related to its exposure during pretraining. We calculated a *Familiarity Index* as:

$$\text{Familiarity Index} = \frac{\ln(1 + \text{annotation count}) + \ln(1 + \text{PMC ID count})}{2},$$

where ln is the natural logarithm, the annotation count is the number of disease annotations for each term and the PMC ID count is the nuber for times the ontology identifier is found in the PMC full-text database. Based on the distribution of the familiarity index (Figure 3), we categorized terms as :

`Unfamiliar`: Index < 3.0
`Somewhat Familiar`: $3.0 \leq$ Index < 4.5
`Highly Familiar`: Index ≥ 4.5

2.8. Hypothesis Testing: Prior Knowledge and Fine-Tuning Outcomes

We hypothesized that a term's prior knowledge state would systematically influence its response to fine-tuning. To test this, we classified outcomes and analyzed prior knowledge effects using a two-stage framework. Each of the 799 test terms was categorized based on deterministic accuracy before and after fine-tuning:

Robustly Correct: Correct both before and after fine-tuning
Gainer: Incorrect before but correct after fine-tuning
Loser: Correct before but incorrect after fine-tuning
Persistent Error: Incorrect both before and after fine-tuning

Stage 1: Prior Knowledge Stratification.
Prior model knowledge for each term was measured along three dimensions using the base model:

Latent probabilistic knowledge: Derived from stochastic accuracy and categorized as Unknown, Weak, Medium, or Strong.
Partial subtoken knowledge: Scored by the number of correct numeric subtokens (None, Weak, Medium, Complete).
Term familiarity: Based on disease annotation frequency and PubMed Central occurrences, categorized as Unfamiliar, Somewhat Familiar, or Highly Familiar.

Stage 2: Fine-Tuning Response Analysis.
We compared deterministic accuracy before and after fine-tuning within each prior knowledge category, using paired t-tests to assess significance and Cohen's d for effect sizes. Bar plots with standard errors visualized group differences, and chi-square tests measured the proportion of terms corrected by fine-tuning relative to baseline. We tested the *reactive middle* hypothesis by examining terms in the middle level of prior knowledge (*Medium* latent probabilistic knowledge, `Medium` partial subtoken

 doi:10.20944/preprints202508.0574.v1

knowledge, and `Somewhat Familiar` term familiarity to determine if they were more likely to be gainers or losers during fine-tuning (Figure 7). Significances of differences in gainers and losers by prior knowledge category were assessed by Chi-square tests.

3. Results

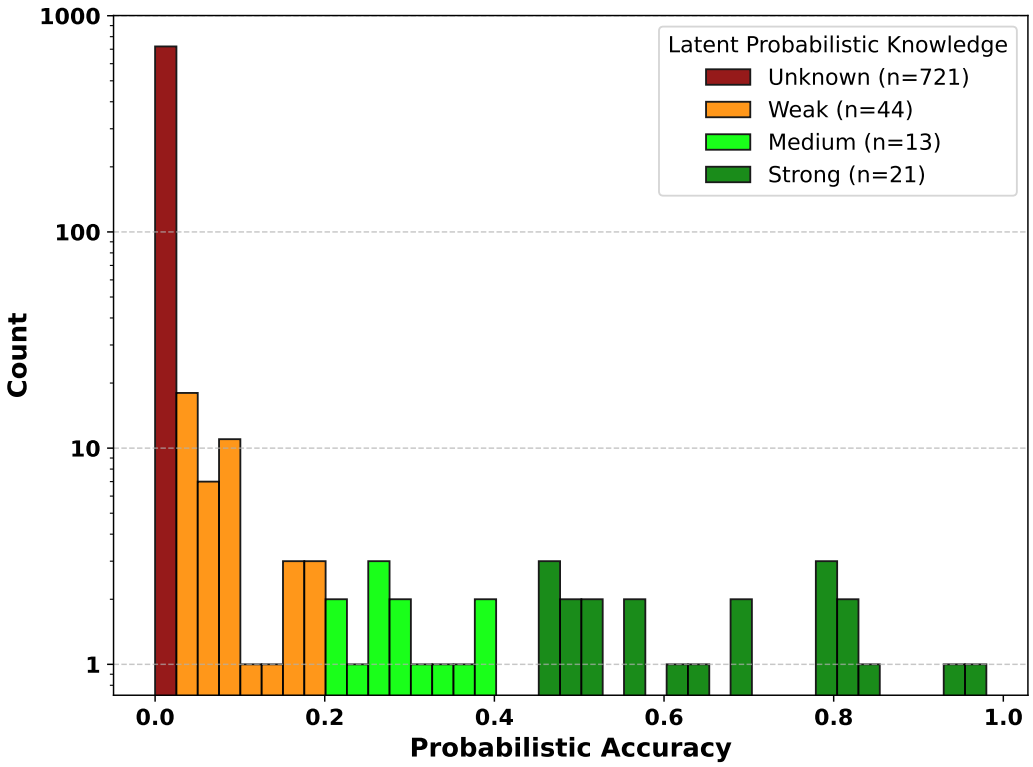

Figure 1. Distribution of latent probabilistic knowledge across 799 HPO terms. Probabilistic accuracy was estimated from 50 stochastic model outputs at temperature 1.0 and binned as `Unknown`, `Weak`, `Medium`, or `Strong`. The y-axis is logarithmic. Most terms were classified as `Unknown`.

 doi:10.20944/preprints202508.0574.v1

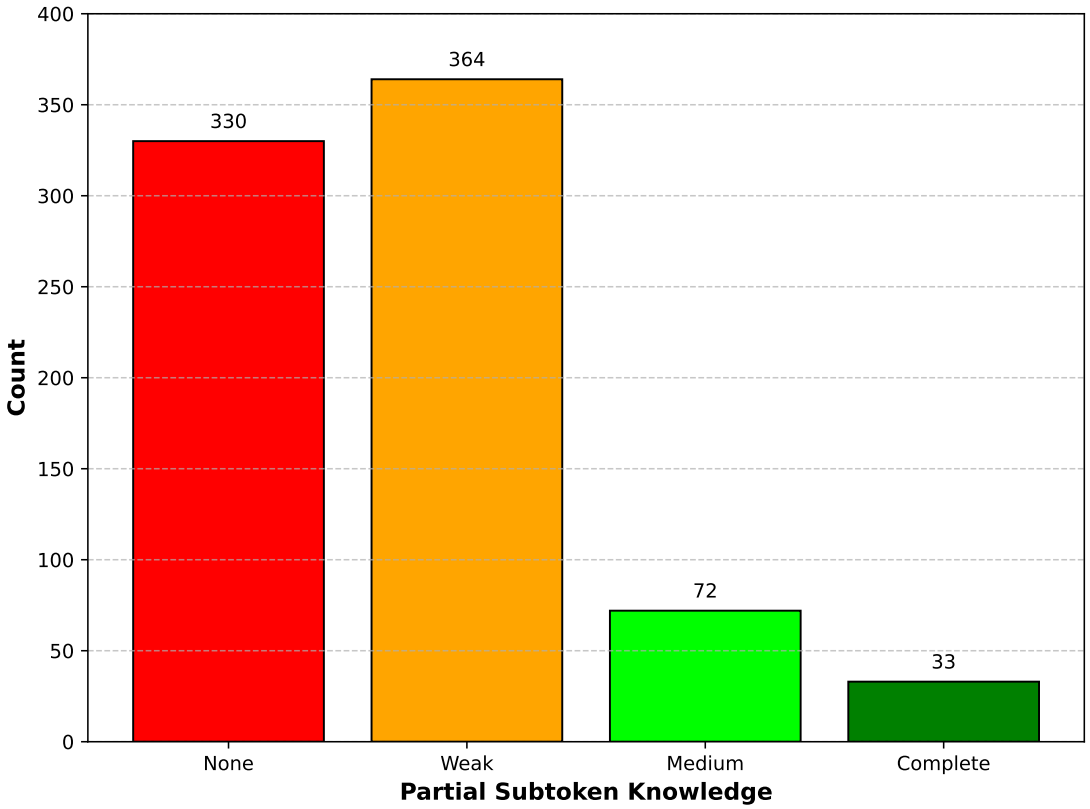

Figure 2. Partial subtoken knowledge categories for predicted HPO identifiers. Bars show counts of terms with `None` (no matching subtokens), `Weak` (1 of 3 subtokens match), `Medium` (2 of 3 match), or `Complete` (all 3 subtokens match). The 7-digit HPO identifier is tokenized into three numeric subtokens by LLaMA 3.1 in the format of 123-456-7.

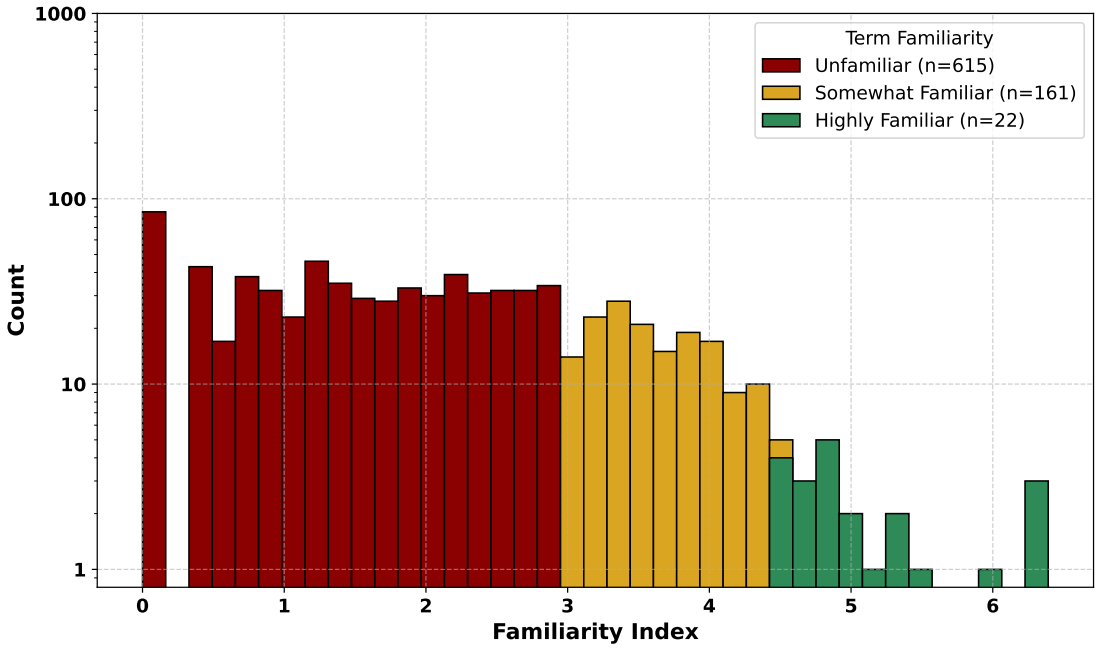

Figure 3. Distribution of term familiarity among 799 HPO terms. Familiarity is calculated from the combined frequency of disease annotations and PubMed Central (PMC) identifier counts. Bins are `Unfamiliar`, `Somewhat Familiar`, and `Highly Familiar`. The y-axis is logarithmic.

 doi:10.20944/preprints202508.0574.v1

3.1. Baseline and Fine-Tuned Performance of LLaMA 3.1 8B on HPO Term Normalization

The baseline LLaMA 3.1 8B model demonstrated significantly higher accuracy on the curated test set of 799 clinically relevant HPO terms compared to the full HPO vocabulary. When evaluated on all 18,988 HPO terms, the model correctly linked only 96 terms (0.5%), whereas on the curated test set, it correctly linked 32 terms (4.0%). A chi-square test confirmed that this difference was highly significant ($\chi^2 = 140.7, p < 1 \times 10^{-30}$). This suggests that the curated test terms represent more common clinical terms that the model was more likely to have encountered during pretraining, making them easier to normalize than the broader, rarer long-tail terms in the complete HPO.

Fine-tuning substantially improved model performance on the curated set of 799 HPO terms. The baseline model correctly linked 32 terms (4.0%), whereas the fine-tuned model correctly linked 118 terms (14.8%). This represents a more than three-fold improvement in deterministic accuracy. A chi-square test confirmed that the increase in correct mappings after fine-tuning was highly significant ($\chi^2 = 58.9, p < 1 \times 10^{-14}$).

3.2. Latent Probabilistic Knowledge Predicts Fine-Tuning Success

Latent probabilistic knowledge, as described in Materials and Methods and adapted from Gekhman et al. [20], was quantified by computing the proportion of correct responses across 50 probabilistic queries (temperature = 1.0). Each term was classified into one of four categories: `Unknown`, `Weak`, `Medium`, or `Strong` (Figure 1).

Figure 4 shows the mean deterministic accuracy of both the base and fine-tuned models for each latent probabilistic knowledge category, with standard error bars and significance markers from two-sample t-tests.

Fine-tuning significantly improved accuracy for terms with `Weak` latent probabilistic knowledge ($p < 0.001$) and modestly for those classified as `Unknown` ($p < 0.001$), though absolute accuracy for `Unknown` terms remained low. For terms with `Medium` latent probabilistic knowledge, accuracy improvements did not reach statistical significance, and for `Strong` latent probabilistic knowledge, both models already performed near ceiling levels with no measurable gain.

These results suggest that fine-tuning is most beneficial for terms where the model possesses partial but incomplete latent probabilistic knowledge, the hypothesized *reactive middle*. Gains are limited for completely unknown terms and negligible for strongly represented terms.

 doi:10.20944/preprints202508.0574.v1

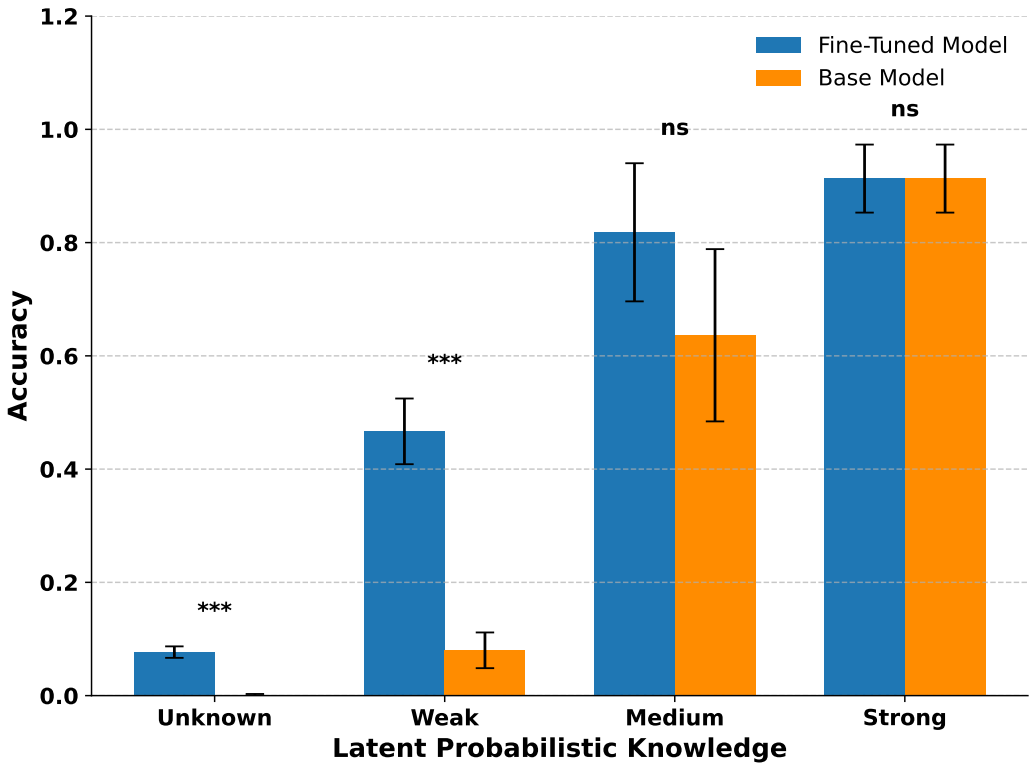

Figure 4. Fine-tuning effects across latent probabilistic knowledge categories. Mean deterministic accuracy of base and fine-tuned models grouped by latent knowledge level (`Unknown`, `Weak`, `Medium`, `Strong`). Fine-tuning yielded the largest gains for `Weak` terms, with smaller improvements for `Unknown`. No significant changes were observed for `Medium` or `Strong` terms. Error bars show standard errors; t-test significance: *** indicates $p < 0.001$, ns indicates not significant.

3.3. Partial Subtoken Knowledge Predicts Fine-Tuning Success

The LLaMA 3.1 tokenizer divides each 7-digit HPO identifier into three numeric subtokens in the format "123–456–7." We assessed partial knowledge by comparing the model's predicted identifier against the ground truth identifier at the subtoken level. Based on the number of correctly predicted numeric subtokens (0–3), each term was assigned to one of four Partial subtoken knowledge categories: `None`, `Weak`, `Medium`, or `Complete`.

Fine-tuning significantly improved deterministic accuracy in the `Weak` and `Medium` categories ($p < 0.001$), but showed no improvement in the `None` category where both models performed poorly. In the `Complete` category, the base model slightly outperformed the fine-tuned model ($p < 0.05$), indicating mild degradation of already well-consolidated knowledge. These results suggest that fine-tuning is most effective when partial knowledge is present but not yet fully established (Figure 5).

 doi:10.20944/preprints202508.0574.v1

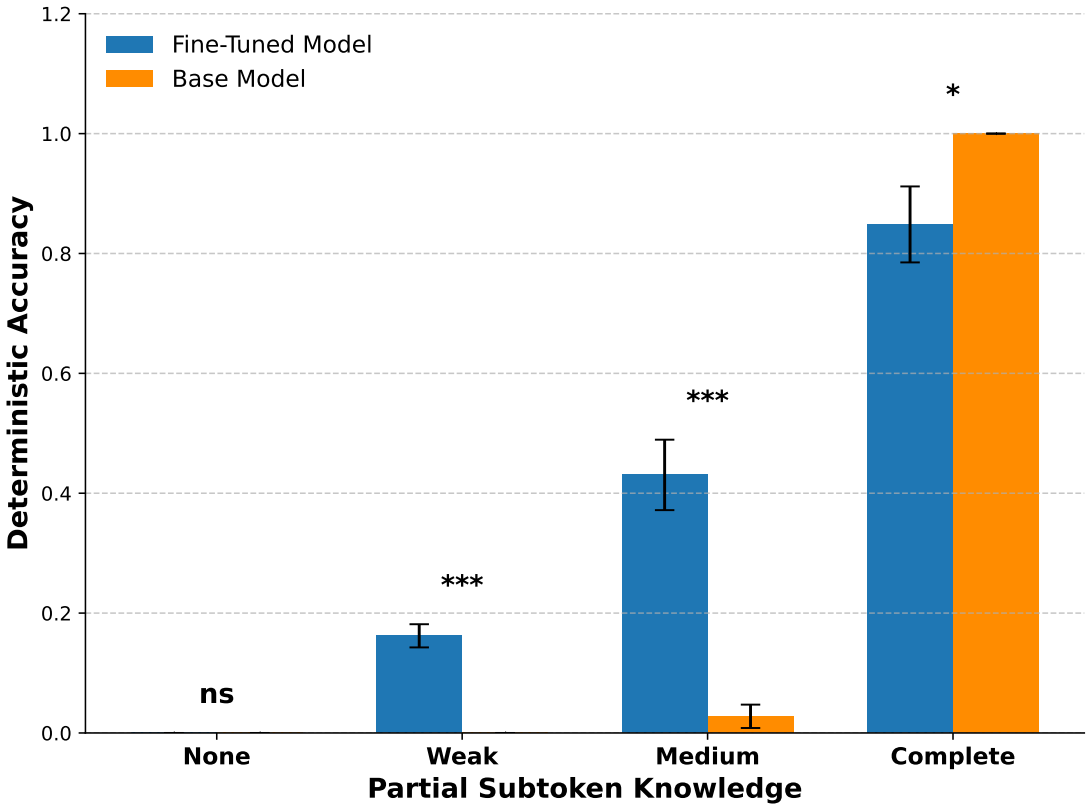

Figure 5. Fine-tuning effects across partial subtoken knowledge categories. Deterministic accuracy for base and fine-tuned models is shown for `None`, `Weak`, `Medium`, and `Complete` subtoken matches. Fine-tuning significantly improved performance in the `Weak` and `Medium` groups but did not improve `None` and slightly degraded accuracy for the `Complete` group. Error bars show standard errors; t-test significance: * indicates $p < 0.05$, *** indicates $p < 0.001$.

3.4. Term familiarity Predicts Fine-Tuning Success

We used the Term familiarity Index, a combined measure of phenotype annotation frequency and PubMed Central (PMC) identifier frequency, as a proxy for prior model exposure to HPO term–identifier pairs during pretraining. Fine-tuning significantly improved deterministic accuracy for both the `Somewhat Familiar` and `Highly Familiar` bins (two-sample t-tests, $p < 0.001$ and $p < 0.01$, respectively). While relative gains were statistically significant for `Unfamiliar` terms, these terms contributed little to the overall increase in correct mappings due to their low absolute accuracy.

These results indicate that terms with moderate to high pretraining exposure account for the majority of fine-tuning gains, whereas rarely seen terms remain challenging to normalize even after fine-tuning (Figure 6).

 doi:10.20944/preprints202508.0574.v1

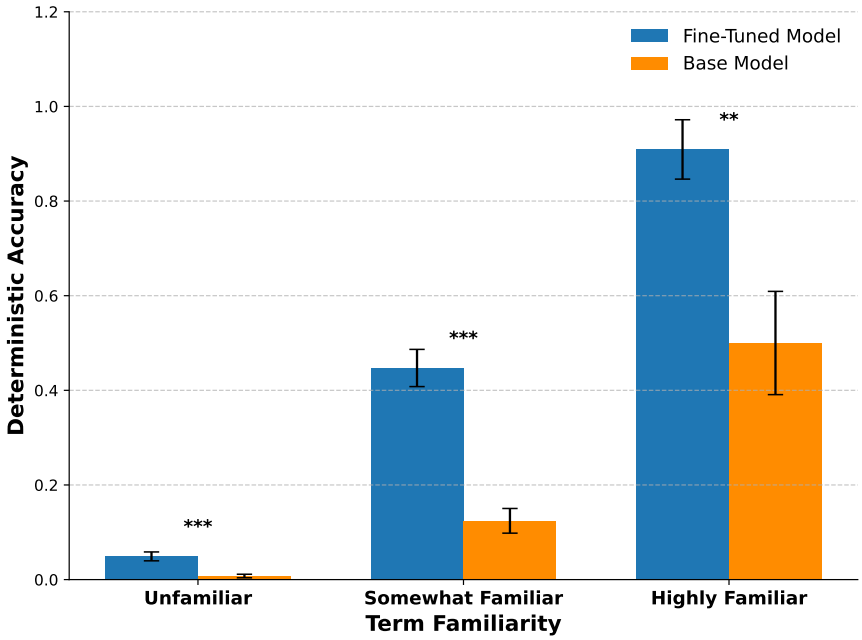

Figure 6. Fine-tuning effects across term familiarity levels. Deterministic accuracy for base and fine-tuned models is shown for `Unfamiliar`, `Somewhat Familiar`, and `Highly Familiar` terms. Fine-tuning produced significant gains for moderately and highly familiar terms. Although `Unfamiliar` terms showed relative improvement, their absolute contribution to accuracy gains was small. Error bars show standard errors; significance: *** indicates $p < 0.001$, ** indicates $p < 0.01$.

3.5. Positive and Negative Knowledge Flows during Fine-Tuning

Pletenev et al. [17] have highlighted that fine-tuning can simultaneously introduce new knowledge and degrade existing knowledge. To quantify these effects, we classified each term as a **Gainer** (incorrect before but correct after fine-tuning) or a **Loser** (correct before but incorrect after). Terms that remained consistently correct or incorrect are not shown in Figure 7.

As illustrated in Figure 7, the majority of gains occurred in the `Unknown` and `Weak` bins of latent probabilistic knowledge, the `Weak` and `Medium` bins of partial subtoken knowledge, and the `Somewhat Familiar` bin of term familiarity. These categories correspond to terms that were neither fully consolidated nor entirely novel, consistent with the reactive middle effect described earlier.

Losses were relatively rare but appeared across multiple bins. In Partial subtoken knowledge, losses occurred most often in the `Complete` bin, indicating that even fully known identifiers can degrade during fine-tuning. Similarly, smaller numbers of losses were observed in the `Medium` and `Strong` bins of Latent probabilistic knowledge and in the `Highly Familiar` bin of Term familiarity, suggesting that well-learned terms are not completely immune to negative knowledge flows.

Overall, fine-tuning produced a net positive knowledge transfer, with gains outnumbering losses in nearly all categories. However, the presence of losses in highly familiar and fully known terms underscores the risk of performance degradation during knowledge injection.

 doi:10.20944/preprints202508.0574.v1

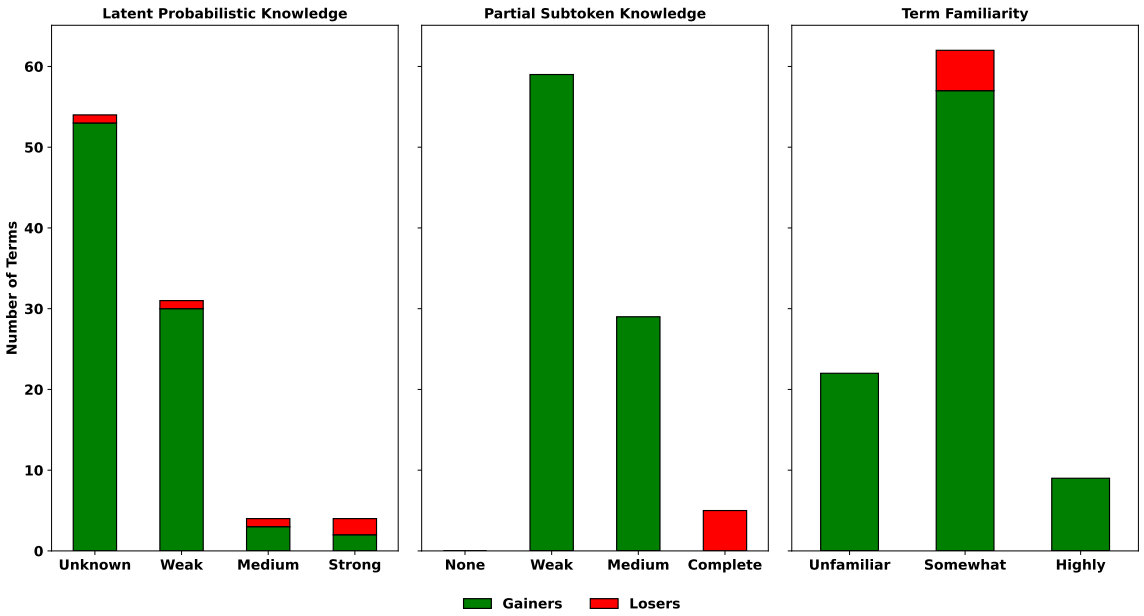

Figure 7. Positive and negative knowledge flows during fine-tuning. Stacked bars show counts of terms that became correct (`Gainers`, green) or became incorrect (`Losers`, red) after fine-tuning. Terms with no change (`Persistent Error` or `Robustly Correct`) are omitted. Chi-square tests indicate significant differences in the distribution of gainers and losers across **Latent Probabilistic Knowledge** ($\chi^2(3) = 20.29$, $p < 0.001$) and **Partial Subtoken Knowledge** ($\chi^2(2) = 93.0$, $p < 0.0001$), but not across **Term Familiarity** ($\chi^2(2) = 2.64$, $p = 0.27$).

4. Discussion

Our findings demonstrate that the success of fine-tuning in improving term-to-identifier linking in biomedical ontologies is systematic rather than random and is strongly influenced by the model's prior knowledge of each term–identifier pair. We evaluated three distinct forms of prior knowledge, latent probabilistic knowledge, partial subtoken knowledge, and term familiarity, and found all three to be predictive of fine-tuning outcomes. Latent probabilistic knowledge refers to hidden or weakly accessible knowledge revealed through stochastic sampling, even when deterministic decoding fails. Partial subtoken knowledge measures the model's ability to reproduce components of an identifier's numeric subtokens correctly, despite failing to generate the complete identifier. Term familiarity captures the degree of prior exposure during pretraining, estimated from annotation frequencies in biomedical databases and identifier frequencies in PubMed Central.

A consistent pattern emerged: terms with intermediate levels of prior knowledge, those neither fully unknown nor well consolidated, were the most responsive to fine-tuning. This reactive middle group exhibited both the largest gains and the most significant losses in deterministic accuracy. Specifically, these terms were characterized by intermediate latent probabilistic knowledge (Figure 1), intermediate partial subtoken knowledge (Figure 2), and moderate term familiarity (Figure 3). In contrast, terms with unknown latent knowledge showed minimal improvement, while highly familiar or strongly known terms tended to remain stable, or, in some cases, experienced slight degradation with fine-tuning.

These observations align with the theoretical frameworks proposed by Gekhman et al.[20] and Pletenev et al.[17], which suggest that fine-tuning can reinforce weakly consolidated representations while simultaneously destabilizing previously stored knowledge. Our findings extend this framework by quantitatively characterizing fine-tuning effects along three axes of prior knowledge, latent probabilistic knowledge, partial subtoken knowledge, and term familiarity. We show that both gains and losses are most pronounced when pre-existing knowledge is partial and unconsolidated. In contrast, when prior knowledge is strongly consolidated (e.g., highly familiar or strongly known terms) or completely absent (unknown terms), the impact of fine-tuning is comparatively small.

 doi:10.20944/preprints202508.0574.v1

We find partial alignment with the results of Wang et al. [28], who demonstrated that parameter-efficient fine-tuning (PEFT) methods such as LoRA can achieve high memorization performance, particularly when trained over many epochs. In their experiments, fine-tuning LLaMA 2 model with 7B parameters over 99 epochs resulted in near-perfect knowledge injection. Unlike our findings, Wang et al. reported no evidence of knowledge degradation or catastrophic forgetting during extensive fine-tuning. We used the LLaMA 3.1 8B model, fine-tuned with Unsloth for only three epochs. This limited training capacity likely contributed to both the modest overall accuracy gains and the performance degradations observed in previously known terms. Unlike Wang et al. [28], we were not able to reach 99 training epochs.

4.1. Limitations

This study has several limitations. First, we evaluated only a single model architecture, LLaMA 3.1 8B, and our findings may not generalize to larger models or those pretrained on domain-specific biomedical corpora. Second, we focused exclusively on the Human Phenotype Ontology (HPO); it remains unknown whether similar patterns would emerge for other biomedical ontologies such as the Gene Ontology (GO), SNOMED CT, or RxNorm. Third, while our ground truth test set of 799 terms was derived from real-world clinical notes, it represents only a small subset of the full HPO vocabulary. Moreover, because the training set included all 18,988 terms while evaluation was limited to 799 terms, a training–test mismatch may have influenced the results. Finally, our fine-tuning procedure was constrained to three epochs with five prompt variants per term. Alternative strategies, including broader prompt diversity, longer training durations, or expanded context windows, may yield different outcomes.

We also did not explicitly address the ongoing debate over whether parameter-efficient fine-tuning (PEFT) methods such as LoRA enable true generalization or merely reinforce memorized associations [15]. Our findings suggest that fine-tuning primarily acts as a structured memorization mechanism, improving mappings for terms already partially known or inconsistently encoded. However, we did not evaluate whether PEFT promotes deeper semantic abstraction beyond the training set.

4.2. Future Work

Future research should investigate whether the observed patterns of fine-tuning responsiveness generalize across other biomedical ontologies, different model sizes, and alternative training strategies. It will be valuable to systematically compare fine-tuning techniques, including QLoRA, Direct Preference Optimization (DPO), and full-model fine-tuning, to assess their effects on knowledge injection and retention.

Additionally, future work should examine how temperature settings influence the estimation of latent probabilistic knowledge, potentially improving our ability to predict fine-tuning outcomes. A particularly promising direction is the exploration of targeted fine-tuning approaches that focus specifically on a *reactive middle*, that is, terms where model prior knowledge is partial and unconsolidated. Such targeted strategies may provide greater efficiency and higher accuracy while minimizing degradation of well-consolidated mappings.

Finally, we plan to study the velocity of fine-tuning, investigating whether certain terms adapt more rapidly to training than others and identifying the predictive factors that govern this responsiveness.

5. Conclusions

This study demonstrates that the effectiveness of fine-tuning for term-to-identifier linking in biomedical ontologies is not uniform but is systematically governed by a model's prior knowledge of term-identifier pairs. We evaluated three distinct dimensions of prior knowledge, latent probabilistic knowledge, partial subtoken knowledge, and term familiarity, and found each to be predictive of fine-tuning outcomes.

 doi:10.20944/preprints202508.0574.v1

Our central finding is the identification of a *reactive middle*: terms with intermediate levels of prior knowledge undergo the most substantial changes during fine-tuning, showing both the largest performance gains and the greatest degradations. In contrast, terms that are either well-consolidated or completely unknown remain largely unaffected by fine-tuning.

These results challenge the assumption that fine-tuning is universally beneficial. Instead, they highlight its dual nature, acting both as a corrective mechanism for incompletely learned mappings and as a potentially disruptive force for fragile but previously accurate knowledge. By explicitly modeling and quantifying prior knowledge, we can better anticipate term-level susceptibility to and vulnerability from fine-tuning, enabling the development of more efficient strategies that maximize accuracy gains while minimizing knowledge degradation.

Author Contributions: Conceptualization, D.B.H. and S.K.P.; Methodology, all authors; Software, A.N., and D.B.H.; Investigation, all authors; Original Draft Preparation, A.N. and D.B.H.; Revisions, all authors; Project Administration, S.K.P.; Funding Acquisition, S.K.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Science Foundation under Award Number 2423235.

Institutional Review Board Statement: The use of EHR clinical notes for research was approved by the Institutional Review Board (IRB) of the University of Illinois (Protocol 2017-0520Z).

Informed Consent Statement: Informed consent was obtained from all participants as part of enrollment in the UIC Neuroimmunology Biobank.

Data Availability Statement: Data and code supporting this study are available from the corresponding author upon reasonable request.

Acknowledgments: We thank the UIC Neuroimmunology Biobank Team for their support.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhou, G.; Zhang, J.; Su, J.; Shen, D.; Tan, C. Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics* **2004**, *20*, 1178–1190.
2. Robinson, P.N. Deep phenotyping for precision medicine. *Human Mutation* **2012**, *33*, 777–780.
3. Köhler, S.; Vasilevsky, N.A.; Engelstad, M.; Foster, E.; McMurry, J.; Aymé, S.; Baynam, G.; Bello, S.M.; Boerkoel, C.F.; Boycott, K.M.; et al. The human phenotype ontology in 2017. *Nucleic Acids Research* **2017**, *45*, D865–D876.
4. Robinson, P.N.; Köhler, S.; Bauer, S.; Seelow, D.; Horn, D.; Mundlos, S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics* **2008**, *83*, 610–615.
5. Jahan, I.; Laskar, M.T.R.; Peng, C.; Huang, J.X. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. *Computers in Biology and Medicine* **2024**, *171*, 108189.
6. Do, T.S.; Hier, D.B.; Obafemi-Ajayi, T. Mapping Biomedical Ontology Terms to IDs: Effect of Domain Prevalence on Prediction Accuracy. In Proceedings of the 2025 IEEE Conference on Artificial Intelligence (CAI). IEEE, 2025, pp. 1–6.
7. Kandpal, N.; Deng, H.; Roberts, A.; Wallace, E.; Raffel, C. Large language models struggle to learn long-tail knowledge. In Proceedings of the International Conference on Machine Learning. PMLR, 2023, pp. 15696–15707.
8. Wu, E.; Wu, K.; Zou, J. FineTuneBench: How well do commercial fine-tuning APIs infuse knowledge into LLMs? *ArXiv* **2024**, *abs/2411.05059*. https://doi.org/10.48550/arXiv.2411.05059.
9. Braga, M. Personalized Large Language Models through Parameter Efficient Fine-Tuning Techniques. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* **2024**. https://doi.org/10.1145/3626772.3657657.
10. Tinn, R.; Cheng, H.; Gu, Y.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; Poon, H. Fine-tuning large neural language models for biomedical natural language processing. *Patterns* **2023**, *4*.

 doi:10.20944/preprints202508.0574.v1

11. Ding, N.; Qin, Y.; Yang, G.; Wei, F.; Yang, Z.; Su, Y.; Hu, S.; Chen, Y.; Chan, C.M.; Chen, W.; et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence* **2023**. https://doi.org/10.1038/s42256-023-00626-4.
12. Wang, C.; Yan, J.; Zhang, W.; Huang, J. Towards Better Parameter-Efficient Fine-Tuning for Large Language Models: A Position Paper, 2023. https://doi.org/10.48550/arXiv.2311.13126.
13. Wu, E.; Wu, K.; Zou, J. Limitations of Learning New and Updated Medical Knowledge with Commercial Fine-Tuning Large Language Models. *NEJM AI* **2025**, p. AIcs2401155.
14. Mecklenburg, N.; Lin, Y.; Li, X.; Holstein, D.; Nunes, L.; Malvar, S.; Silva, B.; Chandra, R.; Aski, V.; Yannam, P.K.R.; et al. Injecting New Knowledge Into Large Language Models Via Supervised Fine-Tuning, 2024.
15. Chu, T.; Zhai, Y.; Yang, J.; Tong, S.; Xie, S.; Schuurmans, D.; Le, Q.V.; Levine, S.; Ma, Y. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161* **2025**.
16. Pan, X.; Hahami, E.; Zhang, Z.; Sompolinsky, H. Memorization and Knowledge Injection in Gated LLMs. *arXiv preprint arXiv:2504.21239* **2025**.
17. Pletenev, S.; Marina, M.; Moskovskiy, D.; Konovalov, V.; Braslavski, P.; Panchenko, A.; Salnikov, M. How Much Knowledge Can You Pack into a LoRA Adapter without Harming LLM? *arXiv (Cornell University)* **2025**. https://doi.org/10.48550/arxiv.2502.14502.
18. Orgad, H.; Toker, M.; Gekhman, Z.; Reichart, R.; Szpektor, I.; Kotek, H.; Belinkov, Y. LLMS Know More Than They Show: On The Intrinsic Representation of LLM Hallucinations, 2025.
19. Gekhman, Z.; Yona, G.; Aharoni, R.; Eyal, M.; Feder, A.; Reichart, R.; Herzig, J. Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations? *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* **2024**, pp. 7765–7784. https://doi.org/10.18653/v1/2024.emnlp-main.444.
20. Gekhman, Z.; David, E.B.; Orgad, H.; Ofek, E.; Belinkov, Y.; Szpektor, I.; Herzig, J.; Reichart, R. Inside-out: Hidden factual knowledge in llms. *arXiv preprint arXiv:2503.15299* **2025**.
21. Amberger, J.S.; Bocchini, C.A.; Schiettecatte, F.; Scott, A.F.; Hamosh, A. OMIM. org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research* **2015**, *43*, D789–D798.
22. Maiella, S.; Rath, A.; Angin, C.; Mousson, F.; Kremp, O. Orphanet and its consortium: where to find expert-validated information on rare diseases. *Revue Neurologique* **2013**, *169*, S3–8.
23. Beck, J.; Sequeira, E. PubMed Central (PMC): An archive for literature from life sciences journals. *The NCBI Handbook* **2003**.
24. Hier, D.B.; Carrithers, M.A.; Platt, S.K.; Nguyen, A.; Giannopoulos, I.; Obafemi-Ajayi, T. Preprocessing of Physician Notes by LLMs Improves Clinical Concept Extraction Without Information Loss. *Information* **2025**, *16*, 446.
25. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. Lora: Low-rank adaptation of large language models. *ICLR* **2022**, *1*, 3.
26. Luo, Y.; Yang, Z.; Meng, F.; Li, Y.; Zhou, J.; Zhang, Y. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747* **2023**.
27. Kalajdzievski, D. Scaling laws for forgetting when fine-tuning large language models. *arXiv preprint arXiv:2401.05605* **2024**.
28. Wang, A.; Liu, C.; Yang, J.; Weng, C. Fine-tuning large language models for rare disease concept normalization. *Journal of the American Medical Informatics Association* **2024**, p. ocae133.