From Memorization to Generalization: Fine-Tuning Large Language Models for Biomedical Term-to-Identifier Normalization

Suswitha Pericharla[1], Daniel B. Hier[2,3], Tayo Obafemi-Ajayi[3]

[1] Computer Science Department, Missouri State University, 901 S. National Avenue, Springfield MO, 65897, MO, USA.
[2]Department of Neurology and Rehabilitation, University of Illinois at Chicago, 912 S. Wood Street, Chicago, 60612, IL, USA.
[3] Engineering Program, Missouri State University, 901 S. National Avenue, Springfield MO, 65897, MO, USA.

Contributing authors: sp5526s@missouristate.edu; dhier@uic.edu; tayoobafemiajayi@missouristate.edu;

Abstract

Background: Effective biomedical data integration depends on automated term normalization—the mapping of natural language biomedical terms to standardized identifiers across terminologies. This linking of terms to identifiers is a crucial step toward semantic interoperability. Although large language models (LLMs) show promise on this task, they exhibit heterogeneous performance across different biomedical terminologies. We evaluated both memorization (performance on training terms) and generalization (performance on unseen validation terms) across three biomedical terminologies: the Human Phenotype Ontology (HPO), Gene Ontology (GO), and protein–gene symbol mappings (GENE).

Results: Fine-tuning Llama 3.1 8B revealed differences across terminologies. GO mappings achieved substantial memorization gains (up to 77% improvement for term→identifier on training terms), whereas fine-tuning yielded minimal gains for HPO. Generalization to unseen validation terms occurred only for protein–gene (GENE) mappings (13.9% improvement), while fine-tuning on HPO and GO produced negligible generalization. Baseline performance varied by model scale, with GPT-4o consistently outperforming both Llama variants for all terminologies. Embedding analyses revealed strong semantic alignment between gene symbols and protein names but no alignment between terms and identifiers for GO or HPO, consistent with differences in lexicalization.

Conclusions: The success of fine-tuning in biomedical term normalization depended on two interacting factors: identifier popularity and identifier lexicalization. Popularity served as a proxy for the likelihood that an LLM encountered a term–identifier pair during pretraining. Despite the high incidence of HPO terms in the literature, the low frequency of corresponding HPO identifiers created a bottleneck that limited memorization. Lexicalization—the degree to which identifiers convey semantic information through embeddings—enabled generalization in the GENE model, where gene symbols behave as if they were natural language tokens. By contrast, the arbitrary identifiers of GO and HPO constrained models to rote memorization. These findings offer a predictive framework for understanding when fine-tuning will enhance factual recall and when it will fail due to low popularity or non-lexicalized identifiers.

Keywords: Large Language Models; Biomedical Ontologies; Fine-tuning; Memorization; Generalization; Term Normalization; Lexicalization; Gene Ontology; Human Phenotype Ontology.

Introduction

The exponential growth of biomedical data has created unprecedented opportunities for advancing healthcare through data mining and machine learning. Realizing this potential requires robust data integration across diverse terminologies and datasets. Large language models (LLMs) have shown promise in addressing these challenges, particularly in clinical text processing and concept extraction. LLMs can process clinical notes and extract medical concepts effectively [1, 2], making them valuable tools for converting unstructured text into computable biomedical data.

Generative autoregressive LLMs, such as models from the GPT and Llama series, produce text by predicting the next token in a sequence. Owing to their large parameter counts and broad training corpora, they exhibit emergent capabilities, notably the ability to retrieve factual knowledge. Although designed for token prediction, they are increasingly being used as repositories of information. This retrieval ability likely reflects two complementary mechanisms: probabilistic token associations learned during pretraining and structured semantic representations encoded in vector embeddings. Wang et al. [3] distinguish between fact-based capabilities (e.g., "What is the HPO identifier for tremor?") and skill-based capabilities (e.g., summarizing a physician note). Biomedical text processing exemplifies an emerging skill domain, but factual retrieval remains a challenge due to both factual gaps and errors (hallucinations) [4].

Precision medicine depends on the exact mapping of biomedical ontology terms to their machine-readable identifiers, ensuring that clinical and biomedical concepts are represented consistently across datasets [5, 6]. This requirement exposes a potential weakness of current LLMs. Even advanced models such as GPT-4o achieve only 8% accuracy when linking Human Phenotype Ontology (HPO) terms to their identifiers [7].

Retrieval-augmented generation (RAG) and supervised fine-tuning have been proposed to address hallucinations and factual gaps [8–11]. Fine-tuning can inject

knowledge [12], but it also carries risks, including knowledge degradation and the overwriting of previously learned facts [13]. Even the best LLMs retrieve only about 65% of the facts they know, and fine-tuning does not always help surface this latent knowledge [14]. Moreover, fine-tuning is most effective for facts that are partially known to the model rather than entirely novel [15, 16]. Fine-tuning performance follows scaling laws so that model size is more important than pretraining data size [17].

Other research highlights additional limitations of fine-tuning, including overreliance on rote memorization, limited generalization to new facts, and the *reversal curse*, which reflects directional biases in knowledge acquisition [3, 14, 18–23]. The balance between memorization and generalization varies systematically across task types: fact retrieval depends more on the memorization, whereas reasoning-intensive tasks rely on generalization during fine-tuning [22].

A key distinction among biomedical terminologies is whether their identifiers are arbitrary codes or lexicalized symbols. As shown in Table 1, terminologies such as HPO and the Gene Ontology (GO) use arbitrary identifiers that are by design opaque codes that are intentionally uninterpretable. In contrast, gene symbols are lexicalized, so that each symbol evokes the name of its referent term, allowing LLMs to leverage preexisting vector embeddings to associate terms with identifiers. When LLMs are fine-tuned on arbitrary codes, they are likely forced to rely on rote memorization of token sequences since these codes provide no intrinsic semantic cues. Lexicalized identifiers, by contrast, can enable a semantic alignment between terms and their machine codes, allowing LLMs to exploit preexisting embeddings rather than depending entirely on the memorization of token sequences. If this account is correct, lexicalized identifiers such as gene symbols [24] would be expected to support some generalization to unseen term–identifier pairs during fine-tuning, whereas arbitrary identifiers would not.

These theoretical distinctions motivated our design of our experiments. We created two fine-tuned models for each of three terminologies—HPO, GO, and gene symbol–protein mappings (referred to as GENE)—and fine-tuned each model bidirectionally (term→identifier and identifier→term). Prior work has shown that RAG-based approaches improve precision for low-prevalence terms [9, 25], and that term popularity in the biomedical literature predicts the ability of LLMs to link terms to identifiers [7, 15]. Building on these insights, we constructed frequency-balanced datasets for all three terminologies to isolate the relative influence of identifier popularity and identifier lexicalization on the success of supervised fine-tuning. Two key patterns emerged:

1. The GO model efficiently memorized term–identifier pairs, whereas the HPO model struggled.
2. The GENE model generalized to unseen term–identifier pairs, while the GO and HPO models failed to do so.

We interpret these findings as reflecting two interacting factors: a *popularity effect*, arising from the long-tail frequency distribution of biomedical terms, and a *lexicalization effect*, reflecting the presence of semantically meaningful identifiers. Arbitrary identifiers such as those found in GO and HPO constrain models to rote memorization, and rare identifiers (as in HPO) further limit memorization efficiency. In contrast,

Table 1: Examples of arbitrary and lexicalized biomedical identifiers.

Terminology	Term	Identifier	Identifier Type
GO	nucleus	GO:0005634	arbitrary
GO	cytosol	GO:0005829	arbitrary
HPO	tremor	HP:0001337	arbitrary
HPO	ataxia	HP:0001251	arbitrary
GENE	fibroblast growth factor	FGF	lexicalized
GENE	breast cancer 1	BRCA1	lexicalized
GENE	tumor protein p53	TP53	lexicalized
GENE	superoxide dismutase 1	SOD1	lexicalized

GO = Gene Ontology (cellular component hierarchy); HPO = Human Phenotype Ontology
GENE = shorthand for protein names linked to gene symbols
Arbitrary identifiers do not resemble the linked term. Lexicalized identifiers are designed to evoke the linked term.

lexicalized identifiers, such as gene symbols, enable semantic alignment that supports limited generalization to unseen mappings. This framework provides systematic predictions of when fine-tuning will succeed or fail in linking biomedical terms to their identifiers.

Methods

We evaluated the ability of large language models to link biomedical terms to their identifiers across three widely used terminologies: HPO [26], GO [27], and UniProtKB protein names [28] with their corresponding HGNC gene symbols [29] (**GENE**). We investigated term–identifier linking in both directions: term → identifier and identifier → term. Figure 1 summarizes the research workflow.

Datasets

HPO contains 18,800 terms describing the phenotypic abnormalities of human diseases. Its primary use is annotating the signs and symptoms of rare disorders, with more than 200,000 annotations that link diseases to their signs [30]. HPO uses arbitrary identifiers.

GO has 43,303 concepts organized into three hierarchies: biological process, molecular function, and cellular component. In this study, we focused on 1,839 terms from the cellular component (CC) hierarchy. GO (CC) provides a standardized terminology for describing the locations of proteins in the cell and adheres to OBO Foundry principles [31]. Across species, GO includes over 9 million annotations [32]. Like HPO, GO employs arbitrary identifiers.

UniProtKB protein names and HGNC gene symbols are interlocking terminologies that link gene symbols, gene names, and protein names. For simplicity, we refer to these interlocking terminologies as **GENE**. The UniProtKB/Swiss-Prot terminology includes approximately 20,000 human proteins, each associated with an HGNC gene symbol [33, 34]. Gene symbols are intentionally chosen to align with their corresponding gene and protein names [35].

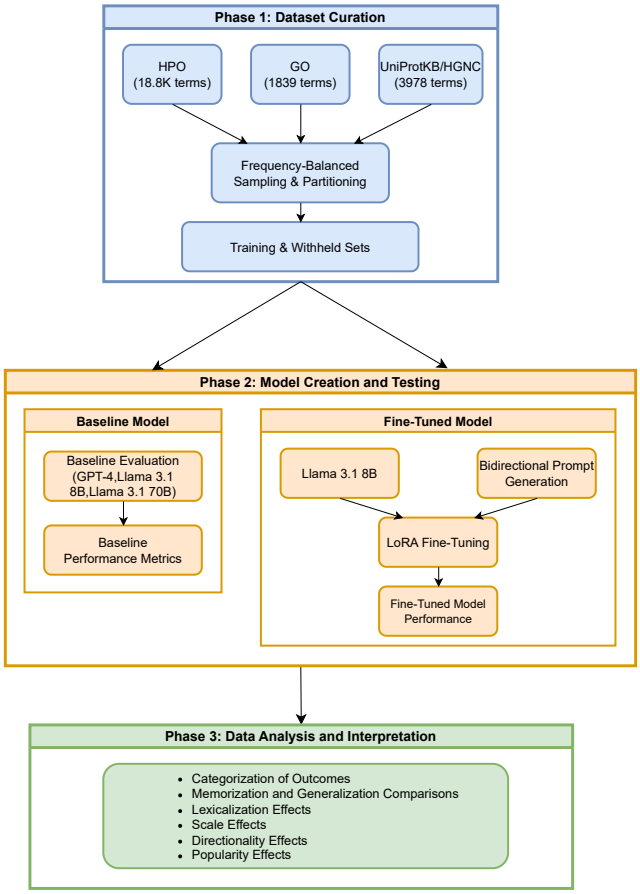

Fig. 1: Overview of Methods. Phase 1: frequency-balanced train and validation test sets were created from three biomedical terminologies. Phase 2: Fine-tuned models were created for each terminology and performance compared to baseline models. Phase 3: Gains from fine-tuning were attributed to either memorization or generalization. Effects of scale, popularity, directionality, and lexicalization were assessed.

The three terminologies differ in structure, size, adoption, and frequency of use in the biomedical literature. HPO and GO are hierarchical ontologies and are distributed in OWL, JSON, and OBO formats via the OBO Foundry [36] and in CSV format via NCBO BioPortal [37]. HPO and GO are organized hierarchically, whereas the UniProtKB/HGNC (GENE) terminologies are flat. Together, they provide a robust testbed for evaluating the memorization and generalization of term–identifier pairs during fine-tuning. Importantly, their contrasting identifier characteristics, arbitrary codes (HPO and GO) versus lexicalized symbols (GENE), provide a direct test of

our hypothesis that lexicalization supports generalization. To ensure a comparable evaluation across terminologies of different sizes and frequencies, we implemented a stratified sampling strategy.

Sampling Strategy

Most biomedical terminologies exhibit a pronounced *long-tail* distribution consistent with Zipf's Law [38]: a small number of head terms occur frequently, while the majority are rare. LLMs often struggle with these long-tail terms, which receive little exposure during pretraining [15, 39–41]. Because it is infeasible to fine-tune on every term, we constructed frequency-balanced datasets for each terminology (HPO, GO, GENE) that included terms sampled from the head, body, and tail of the frequency distribution [7, 42]. Identifier usage counts were retrieved from PubMed Central (PMC) via the PMC API [43]. For each terminology, identifiers were ranked by frequency, partitioned into twenty equal-sized bins, and ten term–identifier pairs were randomly sampled from each bin. Because terms often appear in the biomedical literature without their identifiers, identifier frequency provides a more reliable indicator of long-tail behavior than term frequency alone [15, 42]. This stratified sampling ensured balanced representation across the frequency spectrum and produced 200 pairs for each terminology (HPO, GO, and GENE). These curated datasets formed the foundation for baseline model evaluation before fine-tuning.

Baseline Evaluation

We evaluated baseline performance on unmodified Llama 3.1 models (8B and 70B parameters) [44] and GPT-4o [45]. Each model was tested on the same term–identifier sets using identical prompt templates. This evaluation reflects the intrinsic biomedical knowledge encoded during pretraining and establishes a baseline for subsequent fine-tuning. Performance was assessed in both mapping directions (term $\rightarrow$ identifier and identifier $\rightarrow$ term) across all three datasets (HPO, GO, and GENE). Mapping accuracy was measured using `top-1 accuracy` (hits@1), where a prediction was scored as correct if the model's top-ranked output exactly matched the ground truth identifier (or term). With these baselines established, we next applied parameter-efficient fine-tuning to test whether domain-specific knowledge could be acquired without full model retraining.

Parameter-Efficient Fine-Tuning

For each terminology (HPO, GO, and GENE), equal-sized training sets (200 term–identifier pairs) and validation sets (unseen terms) were constructed. Each training pair was expanded into five distinct prompts to expose the LLM to varied phrasing, as shown in Table 2. Reverse mapping prompts mirrored the forward prompts, substituting identifiers for terms. This produced 1,000 training prompts per terminology in each direction.

We performed fine-tuning on the Llama 3.1 8B Instruct model [46]. Parameter-efficient adaptation used Low-Rank Adaptation (LoRA; rank $r = 64$, scaling $\alpha = 128$)

Table 2: Prompt templates for term–identifier mapping fine-tuning.

Prompt	Template
1	What is the [ONTOLOGY] identifier for the [ONTOLOGY] term [TERM]?
2	The [ONTOLOGY] term [TERM] has what [ONTOLOGY] identifier? Respond only with the [ONTOLOGY] identifier.
3	Provide the [ONTOLOGY] identifier for: [TERM]
4	What is the ontology identifier for [TERM] in [ONTOLOGY]?
5	Return only the [ONTOLOGY] identifier for the term: [TERM]

Each term–identifier pair was expanded into five distinct prompts to capture variation in phrasing. [ONTOLOGY] is the placeholder for the terminology, and [TERM] is the placeholder for the term being probed. Reverse mapping prompts mirrored the forward prompts, substituting identifiers for terms.

[47], applied to all linear transformer layers. Optimization employed supervised fine-tuning with a learning rate of 1×10^{-5} (cosine scheduler, no warm-up, cycle length = 0.5), batch size = 32, max gradient norm = 1, and no weight decay for a total of 20 epochs. The `train-on-inputs` option was set to `auto`. Separate models were trained for each terminology and mapping direction, yielding six fine-tuned models. Validation loss plateaued after ∼5 epochs for HPO and GENE mappings, and ∼15 epochs for GO. We categorized outputs from the fine-tuned models using the following evaluation framework.

Evaluation Framework

Fine-tuning performance was evaluated by comparing accuracy before and after fine-tuning for each test instance (term–identifier or identifier–term pair). Each instance was assigned to one of four mutually exclusive categories: (i) *Gainer*: terms that were *incorrect* at baseline but *correct* after fine-tuning and indicating that knowledge was acquired. (ii) *Loser*: terms that were *correct* at baseline but *incorrect* after fine-tuning indicating that knowledge was degraded. (iii) *Correct*: terms that were *correct* at baseline and after fine-tuning indicating that knowledge was preserved. (iv) *Incorrect*: terms that were *incorrect* at baseline and after fine-tuning. This framework distinguishes gains, losses, and stable outcomes, providing a comprehensive view of fine-tuning effects across terminologies.

We evaluated the performance of the model based on three key dimensions: memorization vs. generalization, lexicalization, and popularity.

Memorization vs. Generalization

Definitions of memorization and generalization in LLM fine-tuning vary across studies [3, 22, 48]. Following Bridgman's notion of *operational definitions* [49], we define these terms strictly by their empirical manifestations.

Memorization: Accuracy gains on mappings from the *training set* (term–identifier pairs explicitly seen during fine-tuning).

Generalization: Accuracy gains on mappings from the *validation set* (term–identifier pairs likely encountered during pretraining but not seen during fine-tuning).

These operational definitions are broadly consistent with recent formal definitions of generalization [50].

Lexicalization

To probe why some terminologies support generalization while others do not, we also examine the role of lexicalization. *Lexicalization* refers here to the degree to which an identifier encodes semantic information that evokes its associated term. Lexicalized identifiers share lexical overlap with the term itself (e.g., `TP53` ↔ `tumor protein p53`), whereas an arbitrary identifier such as `HP:0001250` → `seizure` does not. In this study, we define lexicalization as the presence of semantic alignment between the embeddings of a term and its corresponding identifier. To test for lexicalization, we used Llama 3.1 8B to generate embeddings and then compared the embeddings of terms with those of their identifiers.

We used the 4,096-dimensional embeddings from the Llama 3.1 8B model (`meta-llama/Meta-Llama-3-8B`). Each string (term or identifier) was tokenized, final-layer hidden states were extracted, and token vectors were averaged to yield a single vector. Cosine similarity was computed for each true term–identifier pair and compared with similarities to non-matching identifiers. Higher similarity for matched pairs was interpreted as evidence of lexicalization.

As a complementary measure, all embedding vectors ($n = 600$) were projected into two dimensions using principal component analysis (PCA). Euclidean distances between terms and their paired identifiers were compared with distances to non-paired identifiers. Reduced distances between true pairs were interpreted as additional evidence of lexicalization. This visualization tested whether LLM embeddings organize terms and identifiers within a shared semantic space. Finally, to examine how prior exposure might shape these outcomes, we analyzed proxies for popularity with each term–identifier pair.

Popularity

The frequency with which a fact appears in the training corpus has been termed *popularity* and has been shown to predict both its retrievability and its ease of consolidation during fine-tuning [14, 51]. Facts with higher popularity are easier to recall, whereas rare, long-tail facts are less accessible to the model. In earlier work, we referred to this property as *familiarity* and hypothesized that model familiarity with specific terms and identifiers during pretraining would facilitate fine-tuning success [15]. Here, we adopt the community-standard term *popularity* to emphasize its correspondence with frequency-based effects observed across multiple studies. Because the exact frequencies of facts in pretraining corpora are unknown, we approximate popularity using the following proxies:

1. *Identifier frequency in biomedical text:* counts of HPO identifiers, GO identifiers, and gene symbols in the PubMed Central (PMC) full-text database [43].

Table 3: Baseline and fine-tuned performance across the biomedical terminologies and mapping directions.

Mapping	Llama 8B	Llama 8B FT	Δ FT	Llama 70B	GPT-4o
HPO identifier → Term	0.0	0.0	0.0	0.0	0.2
HPO Term → identifier	0.4	0.6	+0.2	4.7	8.8
GO identifier → Term	0.5	4.0	+3.5	7.2	18.4
GO Term → identifier	2.4	9.8	+7.4	17.5	35.6
Gene → Protein	22.8	32.6	+9.8	51.6	59.7
Protein → Gene	73.7	77.0	+3.3	88.9	95.3

Accuracy values are percentages. Δ FT is the calculated performance gain from fine-tuning Llama 3.1 8B. All models were evaluated on both training and validation terms.

2. *Term frequency in biomedical text:* counts of HPO terms, GO terms, and protein names in PMC.
3. *Annotation frequency in curated resources:* counts of HPO terms used in disease annotation [30], counts of GO cellular component terms used for annotations [32], and counts of protein annotations in the UniProtKB dataset [34].

These three measures serve as operational proxies for popularity, reflecting the likelihood that a model encountered a given term–identifier association during pre-training. Popularity measures (annotation counts, identifier counts in PMC, and term counts in PMC) were Laplace-smoothed by adding one to each value and then $\log_{10}$-transformed to reduce skewness and stabilize variance. The transformed variables were analyzed using two-way analyses of variance (ANOVA) with *Terminology* (HPO, GO, GENE) and *Correctness* (Match = 0 or 1) as fixed factors. When the ANOVA revealed significant main effects, pairwise differences among ontologies were examined using the Games–Howell post hoc test, which does not assume equal variances or sample sizes.

Results

We evaluated baseline performance across three models (Llama 3.1 8B, Llama 3.1 70B, and GPT-4o) on all three terminologies and in both mapping directions. The fine-tuning experiments used the Llama 3.1 8B model exclusively.

Baseline Performance

Table 3 presents baseline accuracy for all models and terminologies. GPT-4o achieved the highest accuracy, followed by Llama 3.1 70B and Llama 3.1 8B. The GENE model showed the highest accuracy (22.8–95.3%), the GO model showed intermediate accuracy (0.5–35.6%), and the HPO model showed the lowest accuracy (0.0–8.8%).

Effect of Fine-Tuning on Accuracy

Fine-tuning effects for Llama 3.1 8B varied by terminology (Table 3). Accuracy gains (ΔFT) were substantial for the GO and GENE models but minimal for the HPO

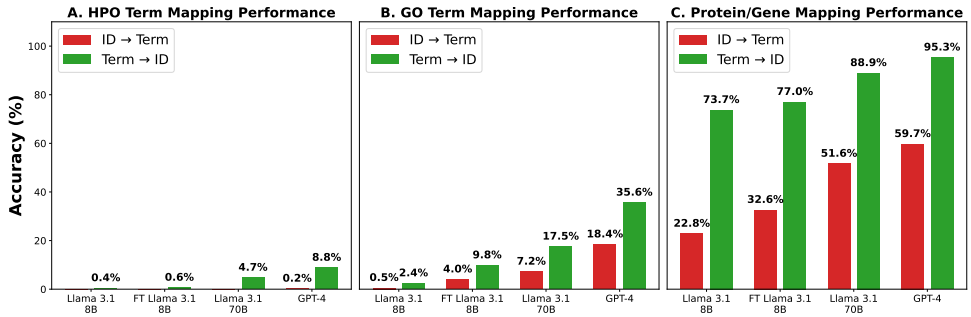

Fig. 2: Bidirectional term mapping performance comparison across biomedical terminologies and model scales. Overall, Term→identifier mappings perform better than identifier→term, with Protein/Gene achieving the highest accuracy, GO moderate, and HPO the poorest across model scales. Performance evaluated on combined training and validation sets.

model. Even after fine-tuning, performance remained below that of the baseline Llama 3.1 70B and GPT-4o models.

Directionality Effects

Figure 2 depicts accuracy by mapping direction. Across all models and terminologies, term → identifier accuracy exceeded identifier → term accuracy, a pattern that held in both baseline and fine-tuned conditions. This directional asymmetry was pronounced across all three systems: GPT-4o achieved 95.3% for protein → gene versus 59.7% for the reverse direction, 35.6% for GO term → identifier versus 18.4% for identifier → term, and 8.8% for HPO term → identifier versus 0.2% for identifier → term.

Knowledge Gains and Losses Following Fine-Tuning

Table 4 classifies each mapping into four mutually exclusive outcomes after fine-tuning: *Gainer*, *Loser*, *Correct*, and *Incorrect*. Results are separated by trained terms (seen during fine-tuning) and validation terms (unseen). These outcomes are also visualized as Sankey flows (Figure 3).

For the HPO model, almost all mappings (97.0–100%) remained *Incorrect*. Gainers comprised only 0.0–2.5% of trained mappings and 0.0–0.3% of validation mappings.

For the GO model, strong memorization occurred on trained terms, with Gainers comprising 33.0–77.0% of trained mappings. In contrast, validation mappings showed minimal gains (0.2–0.7%). GO term → identifier mappings showed 77.0% Gainers in the trained set versus 0.7% in the validation set.

For the GENE model, substantial gains occurred for both trained and validation terms. Gainers comprised 24.0–48.0% of trained mappings and 7.0–13.9% of validation mappings, while Losers comprised 1.5–3.5% of trained mappings and 4.6–6.0% of validation mappings.

Table 4: Assessment of fine-tuning effectiveness on biomedical bidirectional mapping performance.

Terminology	Direction	Category	Validation % (Unseen)	Trained % (Seen)
HPO	identifier → Term	Gainer	0.0	0.0
		Loser	0.0	0.0
		Correct	0.0	0.0
		Incorrect	100.0	100.0
	Term → identifier	Gainer	0.3	2.5
		Loser	0.2	0.0
		Correct	0.3	0.5
		Incorrect	99.3	97.0
GO	identifier → Term	Gainer	0.2	33.0
		Loser	0.4	0.0
		Correct	0.2	0.5
		Incorrect	99.2	66.5
	Term → identifier	Gainer	0.7	77.0
		Loser	1.8	0.0
		Correct	0.6	2.5
		Incorrect	96.8	20.5
GENE	Gene → Protein	Gainer	13.9	48.0
		Loser	6.0	3.5
		Correct	16.7	22.5
		Incorrect	63.4	26.0
	Protein → Gene	Gainer	7.0	24.0
		Loser	4.6	1.5
		Correct	69.2	69.5
		Incorrect	19.2	5.0

Assessment of fine-tuning effectiveness on biomedical bidirectional mapping performance (Training: 200 pairs per terminology; Validation: HPO ∼18,600, GO ∼1,639, GENE ∼3,778 terms). Results are separated into four outcome categories (defined in Methods) and reported for both Trained (Seen) and Validation (Unseen) terms. This breakdown allows comparison of memorization (Seen) and generalization (Unseen) across terminologies.

Derived Metrics of Fine-Tuning Performance

Table 5 summarizes memorization (gains on trained terms), generalization (gains on unseen terms from the validation set), degradation (losses after fine-tuning), and overall accuracy across terminologies.

The HPO model showed virtually no improvement, with minimal memorization (0.0–2.5%) and negligible generalization (0.0–0.3%). The GO model demonstrated strong memorization on trained (seen) terms, reaching up to 77% for term → identifier mappings, but showed minimal generalization (≤1%). In contrast, the GENE model exhibited both robust memorization (24–48%) and notable generalization on the validation set (7–14%), although it was accompanied by some degradation (4–10%).

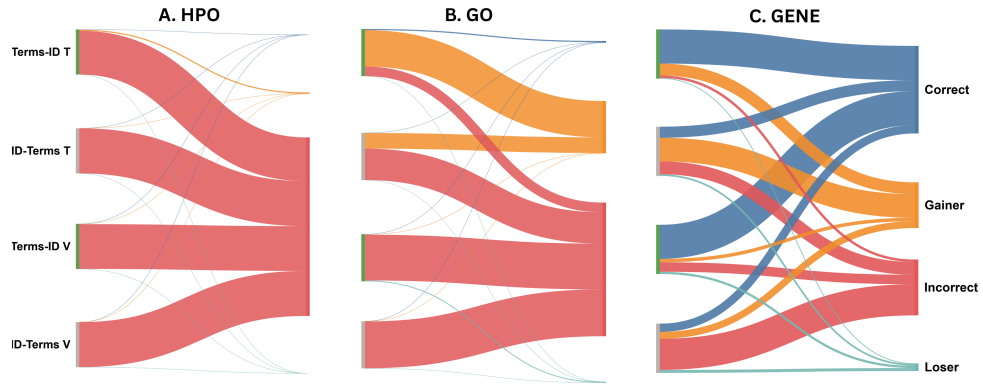

Fig. 3: Sankey diagrams illustrating fine-tuning outcomes across HPO, GO, and GENE models. Color flows visualize how terms transition after fine-tuning: blue = correct, orange = gainers (newly correct), red = incorrect, and green = degraded. Panel A (HPO) shows dominance of red flows, reflecting the long tail of low-popularity identifiers that remain unmapped after fine-tuning. Panel B (GO) exhibits strong orange flows, consistent with moderate popularity and effective memorization of trained terms. Panel C (GENE) displays extensive blue and orange flows, indicating high-popularity, lexicalized identifiers that support both memorization and generalization to unseen terms. Green (degradation) flows are visible mainly in the GENE model, marking the upper limit of representational saturation.

Table 5: Derived performance metrics for fine-tuned Llama 3.1 8B across the three biomedical terminologies.

Task	Memorized	Generalized	Degraded	Accuracy*
HPO identifier → Term	0.0	0.0	0.0	0.0
HPO Term → identifier	2.5	0.3	0.2	3.0
GO identifier → Term	33.0	0.2	0.4	33.5
GO Term → identifier	77.0	0.7	1.8	79.5
Gene → Protein	48.0	13.9	9.5	70.5
Protein → Gene	24.0	7.0	6.1	87.5

All values are percentages. * Accuracy is for trained terms only. Accuracy = Correct + Gainer – Loser
Memorized = gains on trained terms; Generalized = gains on validation terms;
Degraded = losses on trained or validation terms. Values derived from Table 4

Fine-tuning consistently improved performance on trained terms across terminologies, but only the GENE model showed meaningful success on unseen terms.

Lexicalization

PCA projection of 4,096-dimensional embeddings (Figure 4B) showed distinct spatial patterns across terminologies. GENE terms and identifiers formed overlapping clusters. HPO and GO identifiers formed separate clusters distinct from their corresponding

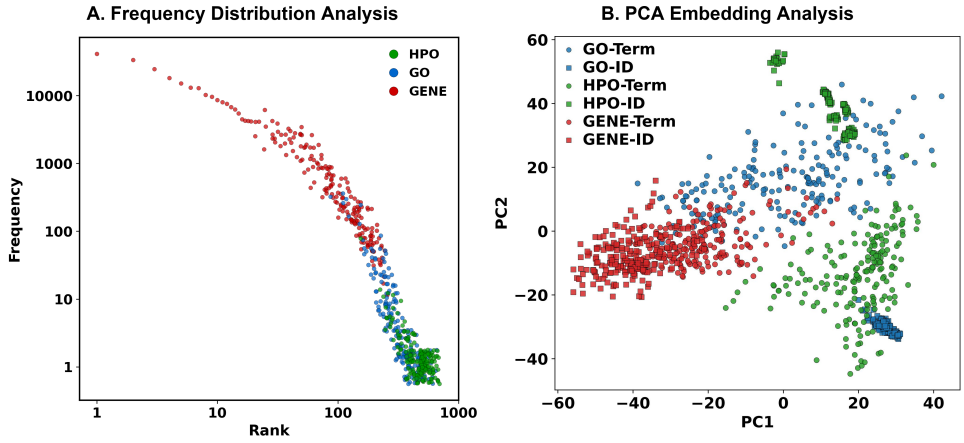

Fig. 4: Identifier frequency and semantic alignment patterns across biomedical terminologies based on the 200 trained terms only. **(A)** Log-log rank-frequency plot of identifiers in PubMed Central showing GENE symbols (red circles) in the distribution head, GO identifiers (blue circles) in the body, and HPO identifiers (green circles) in the long tail. **(B)** PCA projection of Llama 3.1 8B embeddings showing overlapping clusters for GENE terms (red circles) and identifiers (red squares) versus clear separation between terms (circles) and identifiers (squares) for GO (blue) and HPO (green).

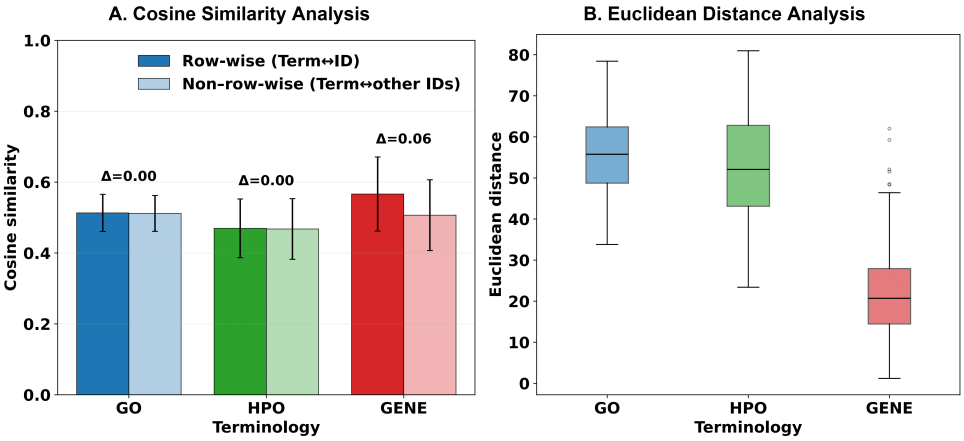

Fig. 5: Semantic alignment between terms and identifiers across biomedical terminologies based on the trained terms only. **(A)** Row-wise vs. non–row-wise cosine similarity reveals semantic alignment only for GENE term–identifier pairs (red bars). Bars show mean cosine similarity (±SD) between each term and its own identifier (solid color) versus the same term paired with all other identifiers in the same terminology (lighter shade). **(B)** Euclidean distances between terms and identifier in PCA space (4096 → 2 dimensions). GENE term–identifier pairs (red box) are close to each other, unlike GO and HPO.

Table 6: Popularity Proxies by Terminology

Measure	Group	Category (N)	Mean	SD
PMC identifier	HPO	Correct (0)	0.0	0.0
		Incorrect (200)	1.1	5.7
	GO	Correct (53)	24.9	84.2
		Incorrect (147)	14.7	40.5
	GENE	Correct (141)	2063.2	5361.0
		Incorrect (59)	1362.5	2717.2
Annotations	HPO	Correct (0)	0.0	0.0
		Incorrect (200)	20.4	133.3
	GO	Correct (53)	715.8	4852.7
		Incorrect (147)	68.3	252.0
	GENE	Correct (141)	37.1	20.7
		Incorrect (59)	31.6	13.9
PMC term counts	HPO	Correct (0)	0.0	0.0
		Incorrect (200)	23 451.7	93 838.4
	GO	Correct (53)	7033.9	35 379.5
		Incorrect (147)	7602.9	45 311.3
	GENE	Correct (141)	2184.1	6109.4
		Incorrect (59)	1094.9	4058.7

Counts are shown in parentheses (N). Means are ± standard deviations (SD).

terms, suggesting that the model does not align textual descriptors with machine-readable identifiers for these ontologies. This analysis tests whether Llama embeddings capture term–identifier correspondence in their internal representational space.

Cosine similarity analysis (Figure 5A) showed higher row-wise similarity (term paired with its own identifier) versus non-row-wise similarity (term paired with other identifiers) for GENE only. GENE row-wise mean: 0.54 (SD = 0.10); non-row-wise mean: 0.48 (SD = 0.08); difference: Δ mean = +0.060; Welch's t = 8.05, $p < 10^{-6}$, n = 200. Although the absolute difference was modest (0.06), it represents a 12.5% relative increase over the baseline mean (0.48), underscoring that the alignment effect is statistically robust and practically meaningful. HPO and GO showed Δ mean = 0.00, indicating no alignment between terms and their identifiers.

Euclidean distance in PCA space (Figure 5 B) confirmed this pattern, showing that GENE term–identifier pairs exhibited smaller distances than HPO or GO pairs.

Popularity

Descriptive statistics for the three proxies of popularity are presented in Table 6. Mean values are shown for PubMed Central (PMC) identifier counts, annotation counts, and PMC term counts, grouped by whether the model correctly or incorrectly performed the term → identifier mapping task. For both PMC identifier counts and annotation counts, correctly mapped terms exhibited higher values—indicating greater popularity—than incorrectly mapped terms. This relationship was less consistent for PMC

term counts, suggesting that raw term frequency alone does not fully capture model familiarity.

Among the three ontologies, PMC identifier counts differed sharply, following a clear ordering of **GENE** > **GO** > **HPO**. The mean identifier frequency in PMC was approximately 200-fold higher for GENE identifiers (gene symbols) than for HPO identifiers, with GO occupying an intermediate position. This pattern was confirmed by two-way ANOVA and Games–Howell post hoc tests, which showed significant pairwise differences among all three ontologies (all $p < 0.001$). The strong gradient in PMC identifier counts indicates that models were likely exposed far more frequently to gene symbols than to GO or HPO identifiers during pretraining—consistent with the hypothesis that greater corpus exposure enhances representational strength and retrieval accuracy.

Discussion

This study evaluated the ability of large language models to learn biomedical term–identifier mappings across three representative terminologies (HPO, GO, and GENE) and three LLM models (Llama 3.1 models (8B and 70B parameters) and GPT-4o . Several consistent patterns emerged from the results. First, model scale effects prior to fine-tuning were evident: larger models consistently outperformed smaller ones. Second, directionality effects were robust: mapping terms to identifiers was more accurate than mapping identifiers to terms across all models and terminologies. Third, the effectiveness of fine-tuning varied across terminologies: fine-tuning substantially improved GO and GENE mappings, but yielded negligible gains for HPO. Finally, only the GENE terminology demonstrated evidence of generalization, in which the model surfaced unseen term–identifier pairs after fine-tuning. By contrast, improvements for HPO and GO were restricted to memorization of training examples. We interpret these findings in light of four explanatory factors: model scale, directionality, popularity, and lexicalization.

Model Scale Effects

Across all tasks, larger models outperformed smaller ones, with GPT-4o achieving the highest baseline accuracy, followed by Llama 3.1 70B and Llama 3.1 8B (Table 3). Fine-tuning improved the 8B model relative to its own baseline, but it never surpassed the performance of the larger pretrained models. This pattern reflects the well-documented scaling laws of LLMs [17, 52, 53], where greater parameter counts, larger embedding dimensions, and more extensive pretraining exposure yield stronger performance. Even with parameter-efficient fine-tuning, architectural capacity and pretraining scale remain the dominant determinants of success.

Directionality Effects

Another robust pattern was the consistent directionality effect: mapping terms to identifiers was always more accurate than mapping identifiers to terms, regardless of model

size or terminology (Figure 2). We hypothesize that this asymmetry is an inherent artifact of the autoregressive training objective of LLMs. During pretraining, models are optimized to predict the next token given a preceding context. In biomedical corpora, natural language terms usually precede their identifiers (e.g., *ataxia* → HP:0001251). Thus, the forward direction (term → identifier) aligns with the statistical structure of pretraining data, while the reverse direction does not.

This explanation is consistent with prior observations of order sensitivity in LLM reasoning tasks, including the so-called "reversal curse" [19], where models trained on "A is B" fail to infer the logically equivalent "B is A." Applied here, the phenomenon suggests that identifiers are not encoded as flexible semantic units, but rather as fixed token sequences whose predictability depends strongly on their position relative to terms. We next examine how popularity and frequency distributions across terminologies influence the success or failure of fine-tuning.

Popularity Effects

The results identified an additional explanatory factor is popularity, defined as the frequency with which terms and identifiers appear in biomedical text corpora (e.g., PubMed Central). Popularity reflects the likelihood that a fact will be encountered by the model during pretraining. Fine-tuning success was strongly correlated with identifier frequency: GO identifiers appeared more often than HPO identifiers, while gene symbols were orders of magnitude more frequent still (Table 6).

This frequency effect parallels what recent theoretical work describes as *factual salience*—the internal strength with which facts are encoded in model weights after pretraining [54]. In this sense, factual salience can be viewed as a manifestation of the popularity of a fact within the model's parameters. Table 6 quantifies these patterns: GENE identifiers appeared approximately 200-fold more frequently in PMC than HPO identifiers, directly corresponding to the observed differences in fine-tuning success rates. Ghosal et al. [54] demonstrated that the number of times a fact is seen during pretraining correlates with its salience, implying that corpus-level popularity serves as a practical proxy for how strongly knowledge is stored. Highly salient facts—those with strong pretrained associations between a term and its identifier—are more easily retrieved and require fewer fine-tuning updates for consolidation. Conversely, low-salience facts, typically drawn from the long tail of rare identifiers, pass through a prolonged *guessing phase* before stable memorization is achieved.

In a complementary view, Gekhman et al. [20, 55] propose that fine-tuning is most effective for partially known or latent facts that are already weakly represented in the model, whereas entirely unfamiliar facts are harder to integrate and may even promote hallucination. Similarly, Chang et al. [51] showed that popularity mediates the transition from memorization to generalization: as exposure increases, models shift from rote token-level learning to semantically grounded retrieval. Together, these perspectives suggest that popularity not only strengthens factual salience but also fosters the formation of semantic embeddings that enable generalization. Within this framework, popularity in text corpora can be viewed as an observable correlate of both factual salience and latent knowledge within the model's representational space.

The distributional skew of this effect is evident in the Zipf-style log–log rank–frequency plot (Figure 4 A), where gene symbols dominate the head of the distribution, GO identifiers occupy the mid-body, and HPO identifiers are concentrated in the long tail—consistent with the frequency gradient GENE > GO > HPO. This long-tail structure helps explain why fine-tuning yielded negligible gains for HPO: most sampled terms were rare, limiting the model's ability to benefit from fine-tuning (Table 5). This interpretation aligns with the view of Ghosal et al. [54], who describe low-frequency facts as having low factual salience, and of Gekhman et al. [20, 55], who argue that fine-tuning is most effective for partially known (latent) rather than entirely unfamiliar facts. Together, these perspectives situate popularity within a broader theoretical framework linking corpus frequency, knowledge salience, and fine-tuning efficiency.

Surprisingly, even though GO had lower PMC identifier counts than GENE, memorization results were greater for GO than for GENE. This finding aligns with our prior results [15], suggesting that fine-tuning is most effective in the *reactive middle* of the term-frequency distribution. Fine-tuning is inefficient for very rare (tail) terms, which lack sufficient prior exposure for consolidation, and for very common (head) terms, which are already well encoded and exhibit ceiling effects [15]. Because most GENE terms occupy the high-frequency end of the distribution (Figure 4 A), they display limited additional gains from fine-tuning compared with GO terms.

Other studies (e.g., Wang et al. [3]) have shown that brute-force strategies (up to 100 training epochs) can improve performance on difficult-to-memorize HPO terms from the long tail, albeit inefficiently. Taken together, these results suggest that popularity—and its representational analog, factual salience—acts as a strong predictor of memorization success during fine-tuning until a ceiling is reached, where gains plateau due to extreme popularity. Nonetheless, popularity alone does not explain why the GENE model successfully mapped unseen term–identifier pairs from the validation set. To account for this, we next examine lexicalization.

Lexicalization Effects

While popularity predicts the ease of memorization, only lexicalization explains why the GENE model generalized to unseen term–identifier pairs (Table 4, 5). The outcome categories (Gainer, Loser, Correct, Incorrect) reveal that GENE achieved gains on both trained and validation terms, whereas GO gains were restricted to trained terms and HPO showed minimal gains overall (Table 4, Figure 3). We use *generalization* operationally to mean that the model correctly retrieved term–identifier pairs after fine-tuning that were not seen during training. The GENE model's ability to map unseen protein–gene symbol pairs is best explained by the lexicalization of gene symbols, which provides semantic cues absent from the arbitrary codes of GO and HPO.

The term *lexicalization* originated in traditional linguistics and refers to the process by which symbolic expressions or acronyms (e.g., radar, laser) become conventionalized as words in natural language [56]. Even the term "LLM" is a contemporary example—it has undergone rapid lexicalization, evolving from an acronym to a widely recognized word in less than a decade. In the context of large language models, we

extend this notion to describe how symbolic identifiers acquire semantic embeddings during pretraining that position them near meaningful words in vector space. We posit that an identifier is *lexicalized* when its embedding aligns with that of its corresponding natural-language term, allowing the model to rely on semantic rather than purely sequential associations.

Our embedding analyses support this view. Cosine similarities based on Llama 3.1 embeddings (Figure 5 A) showed that protein–gene pairs exhibited significantly higher within-pair similarity than random comparisons, indicating above-chance semantic alignment. PCA visualizations (Figure 4 B) reinforced this finding: gene symbols intermingled with protein names in embedding space, whereas HPO and GO identifiers remained clustered apart from their terms. Euclidean distance analyses (Figure 5 B) further confirmed that GENE term–identifier pairs were located much closer together than HPO or GO pairs.

Taken together, these results suggest that lexicalization provides a representational mechanism for generalization. When identifiers are arbitrary codes (HPO, GO), fine-tuning reinforces memorized token sequences. When identifiers are lexicalized (GENE), fine-tuning leverages semantic overlap to extend learning to unseen terms. Thus, lexicalization helps explain why generalization was observed exclusively in the fine-tuned GENE model. Our interpretation of generalization is closely aligned with recent work demonstrating that fine-tuning on even a small subset of memorized associations with semantically meaningful prompts can "unlock" broader latent knowledge already present in the model's parameters [50]. Once the model aligns a few examples with specific semantic cues, it can retrieve other facts encoded under the same relational structure — even if those specific instances were never presented during fine-tuning. This perspective matches our observation that correct predictions on a subset of protein–gene symbol pairs led to generalization across additional, unseen pairs, suggesting that fine-tuning acts less as a source of new knowledge and more as a mechanism for activating existing representations. In this sense, our results empirically support the theoretical framework of Wu et al. [50], demonstrating that the same principle holds in the biomedical normalization domain, where lexicalized identifiers enable latent associations to be activated beyond the fine-tuned examples.

Importantly, these gains did not come without cost. As noted by Pletenev et al. [13], fine-tuning can degrade existing knowledge, and in our experiments, losses were greatest for the GENE model (Table 5, column *Degraded*). This dual outcome—generalization coupled with degradation—underscores the need to consider popularity and lexicalization as interacting influences during fine-tuning.

How Popularity Holds the Key to Lexicalization

Popularity and lexicalization are best understood as interacting forces rather than independent explanations. At low levels of popularity, identifiers remain arbitrary, and fine-tuning yields little benefit (as with HPO). With moderate popularity (GO), repeated exposure strengthens token-level associations, allowing fine-tuning to support rote memorization. However, once popularity exceeds a critical threshold, a qualitative shift occurs: identifiers begin to acquire meaning through their co-occurrence

patterns, and the model transitions from mere sequence memorization to embedding-based representation. This transition resembles a tipping point rather than a smooth continuum—an abrupt reorganization of the model's representational space rather than a gradual progression.

Crucially, lexicalization is not created by fine-tuning but is inherited from pre-training. During pretraining, high-frequency identifiers repeatedly co-occur with their descriptive terms, allowing their embeddings to converge in semantic space [57]. Fine-tuning then exploits this pre-established alignment by adjusting decision boundaries to make latent associations accessible. This interpretation aligns with Chang et al. [51], who demonstrated that popular facts are not only more likely to be memorized but also more likely to develop semantically coherent embeddings that support generalization. In this view, popularity drives a representational phase change: repetition transforms statistical co-occurrence into stable, vectorized alignments that constitute lexicalization. Moreover, this tipping point or phase shift likely occurs during pretraining, not during fine-tuning.

Accordingly, rare mappings (unpopular facts) are difficult to memorize; moderately frequent mappings can be memorized but not generalized; and highly frequent mappings cross the lexicalization threshold, acquiring embeddings that support semantic alignment and generalization. Once this threshold is crossed, identifiers cease to be arbitrary tokens and function instead as meaningful lexical units. Generalization to unseen term–identifier pairs after fine-tuning thus emerges not from incremental exposure, but from a paradigmatic shift in how the model encodes meaning—one that reorganizes representation rather than merely reinforcing memorized sequences.

Limitations

While our findings provide new insights into the mechanisms of fine-tuning success, we do acknowledge that there are some limitations. Note that we attributed all gains on training terms to memorization, although some may reflect embedding-based alignment. This yields a conservative estimate of generalization. We defined generalization as the ability of the fine-tuned model to learn unseen term–identifier pairs from the validation set, recognizing that the precise definition of generalization for factual fine-tuning remains unsettled. In addition, the balancing of test sets across head, body, and tail frequency distributions relied on PubMed Central identifier counts and ontology annotation counts as surrogates for pre-training exposure. Because true pre-training frequencies are unknown, these proxies are imperfect. The strong long-tail structure of HPO meant that most sampled terms were rare, partly explaining its low memorization rates (Table 5).

The mapping task itself—linking natural-language terms to arbitrary identifiers—represents a special case of fact memorization. This asymmetry applies to HPO and GO but not to protein–gene symbol pairs, and results may not generalize to symmetric fact pairs where both elements carry semantic content (e.g., Paris $\leftrightarrow$ France). For this work, we analyzed embeddings only in the base Llama 3.1 8B model, assuming that fine-tuning re-weights token probabilities more than it reshapes embedding geometry. This assumption remains untested. We also used a single PEFT setup (LoRA on Llama 3.1 8B) with fixed hyper-parameters. Alternative adapters, data-mixing

strategies, or larger base models may alter the memorization–generalization trade-off. We also attributed knowledge loss (degradation) to parameter overwriting, though the mechanisms are likely related to catastrophic forgetting. Finally, although lexicalized identifiers appear to support generalization, the underlying mechanism remains uncertain. Whether fine-tuning primarily exploits pre-existing embedding structures or induces new representational dynamics is an open question. Moreover, this study does not address recent proposals that factual associations may be stored in localized *knowledge neurons* or *knowledge subnetworks* within transformer architectures [58, 59].

Future Work

Future studies should extend these findings in several ways. First, larger and more comprehensive test sets are needed to confirm robustness. Expanding sampling across the full frequency range of ontology terms would strengthen evidence for the roles of popularity and lexicalization. Applying the same framework to additional terminologies, such as ICD-10, SNOMED CT, RxNorm and LOINC, would test whether the observed continuum from memorization to generalization generalizes across biomedical vocabularies. In addition, evaluating larger base models (e.g., Llama 3.1 70B) and other foundation models (e.g., GPT-4o, Mistral, Claude) would clarify how scale and architecture influence the memorization–generalization balance. Systematic exploration of fine-tuning parameters—adapter types, learning-rate schedules, and data-mixing strategies—could reveal methods that reduce knowledge degradation while preserving generalization. Finally, direct analysis of embeddings in fine-tuned models may determine whether fine-tuning primarily re-weights existing representations or reshapes semantic structure.

Conclusion

This study evaluated the capacity of large language models to normalize biomedical terms across three contrasting terminologies: the Human Phenotype Ontology (HPO), the Gene Ontology (GO), and protein–gene symbol mappings (GENE). Using Llama 3.1 (8B and 70B) and GPT-4o as base models, and fine-tuning the Llama 3.1 8B model, several consistent patterns emerged. The findings reveal that popularity—defined as exposure to term–identifier pairs during pretraining—proved essential for memorization. Large language models leverage popularity to consolidate *factual salience*, the internal strength with which facts are encoded in model parameters. GO identifiers, which occur more frequently in biomedical text than HPO identifiers, supported robust memorization after fine-tuning, whereas the long-tailed, infrequent HPO identifiers showed only modest gains. In addition, *lexicalization* emerged as the key to generalization. Protein names and gene symbols are semantically aligned, allowing the model to retrieve term–identifier pairs not seen during fine-tuning but already represented in model parameters through pretraining. By contrast, the arbitrary identifiers of GO and HPO constrained the model to rote memorization. Lexicalization appears to arise during pretraining, when highly frequent identifiers (high popularity) acquire meaningful embeddings (high factual salience).

Directionality and scale effects were consistent. Across all terminologies, mappings from *term→identifier* uniformly outperformed *identifier→term*, reflecting the autoregressive bias of large language models. Larger models consistently outperformed smaller ones, and fine-tuning the 8B model did not close the gap with Llama 70B or GPT-4o. Knowledge gains came with trade-offs. Fine-tuning improved performance on both trained and some unseen terms but also degraded portions of preexisting knowledge—particularly for GENE mappings, where baseline accuracy was already high.

Together, these findings indicate that fine-tuning promotes memorization when identifiers have high popularity in the training corpus, whereas generalization emerges only when identifiers are lexicalized. This *popularity–lexicalization continuum* provides a predictive framework for understanding when fine-tuning will succeed, when it will fail, and why some terminologies remain resistant to factual knowledge acquisition. Crossing the lexicalization threshold is not merely a gradual gain in memorization capacity, but a transition from meaning encoded in token patterns to meaning residing in semantically aligned structures.

References

[1] Hier DB, Carrithers MD, Platt SK, Nguyen A, Giannopoulos I, Obafemi-Ajayi T. Preprocessing of Physician Notes by LLMs Improves Clinical Concept Extraction Without Information Loss. Information. 2025;16(6):446. https://doi.org/10.3390/info16060446.

[2] Hier DB, Munzir SI, Stahlfeld A, Obafemi-Ajayi T, Carrithers MD. High-Throughput Phenotyping of Clinical Text Using Large Language Models. In: 2024 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI). IEEE; 2024. p. 1–4.

[3] Wang Z, Zhao S, Wang Y, Huang H, Xie S, Zhang Y, et al. Re-task: Revisiting LLM tasks from capability, skill, and knowledge perspectives. arXiv preprint arXiv:240806904. 2024;.

[4] Zhang Y, Li Y, Cui L, Cai D, Liu L, Fu T, et al. Siren's song in the AI ocean: a survey on hallucination in large language models. arXiv preprint arXiv:230901219. 2023;.

[5] Rayan RA, Zafar I. Precision Medicine in the Context of Ontology. Semantic Web for Effective Healthcare. 2021;p. 191–213.

[6] Robinson PN, Mungall CJ, Haendel M. Capturing phenotypes for precision medicine. Molecular Case Studies. 2015;1(1):a000372.

[7] Do TS, Hier DB, Obafemi-Ajayi T; IEEE. Mapping Biomedical Ontology Terms to IDs: Effect of Domain Prevalence on Prediction Accuracy. 2025 IEEE Conference on Artificial Intelligence (CAI). 2025;p. 1–6.

[8] Wang A, Liu C, Yang J, Weng C. Fine-tuning Large Language Models for Rare Disease Concept Normalization. bioRxiv. 2023;p. 2023–12.

[9] Hier DB, Do TS, Obafemi-Ajayi T. A simplified retriever to improve accuracy of phenotype normalizations by large language models. Frontiers in Digital Health. 2025;7:1495040.

[10] Shlyk D, Groza T, Mesiti M, Montanelli S, Cavalleri E. REAL: A Retrieval-Augmented Entity Linking Approach for Biomedical Concept Recognition. In: Proceedings of the 23rd Workshop on Biomedical Natural Language Processing; 2024. p. 380–389.

[11] Thompson WE, Vidmar DM, De Freitas JK, Pfeifer JM, Fornwalt BK, Chen R, et al. Large Language Models with Retrieval-Augmented Generation for Zero-Shot Disease Phenotyping. arXiv preprint arXiv:231206457. 2023;.

[12] Keloth VK, Hu Y, Xie Q, Peng X, Wang Y, Zheng A, et al. Advancing entity recognition in biomedicine via instruction tuning of large language models. Bioinformatics. 2024;40(4):btae163.

[13] Pletenev S, Marina M, Moskovskiy D, Konovalov V, Braslavski P, Panchenko A, et al. How Much Knowledge Can You Pack into a LoRA Adapter without Harming LLM? arXiv preprint arXiv:250214502. 2025;.

[14] Yuan J, Pan L, Hang CW, Guo J, Jiang J, Min B, et al. Towards a holistic evaluation of LLMs on factual knowledge recall. arXiv preprint arXiv:240416164. 2024;.

[15] Hier DB, Platt SK, Nguyen A. Prior Knowledge Shapes Success When Large Language Models Are Fine-Tuned for Biomedical Term Normalization. Information. 2025;16(9):776.

[16] Speicher T, Khan MA, Wu Q, Nanda V, Das S, Ghosh B, et al. Understanding memorisation in llms: Dynamics, influencing factors, and implications. arXiv preprint arXiv:240719262. 2024;.

[17] Zhang B, Liu Z, Cherry C, Firat O. When Scaling Meets LLM Finetuning: The Effect of Data, Model and Finetuning Method. arXiv preprint arXiv:240217193. 2024;.

[18] Ovadia O, Brief M, Mishaeli M, Elisha O. Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs. arXiv preprint arXiv:231205934. 2024;.

[19] Berglund L, Tong M, Kaufmann M, Balesni M, Stickland AC, Korbak T, et al. The Reversal Curse: LLMs trained on A is B fail to learn B is A. arXiv preprint arXiv:230912288. 2023;.

[20] Gekhman Z, Yona G, Aharoni R, Eyal M, Feder A, Reichart R, et al. Does fine-tuning LLMs on new knowledge encourage hallucinations? arXiv preprint arXiv:240505904. 2024;.

[21] Li B, Liang H, Li Y, Fu F, Yin H, He C, et al. Gradual Learning: Optimizing Fine-Tuning with Partially Mastered Knowledge in Large Language Models. arXiv preprint arXiv:241005802. 2024;.

[22] Wang X, Antoniades A, Elazar Y, Amayuelas A, Albalak A, Zhang K, et al. Generalization v.s. Memorization: Tracing Language Models' Capabilities Back to Pretraining Data. In: International Conference on Learning Representations; 2025. Available from: https://openreview.net/forum?id=RdJVFCHjUMI.

[23] Wu E, Wu K, Zou J. FineTuneBench: How well do commercial fine-tuning APIs infuse knowledge into LLMs? arXiv preprint arXiv:241105059. 2024;.

[24] Bruford EA, Braschi B, Denny P, Jones TE, Seal RL, Tweedie S. Guidelines for human gene nomenclature. Nature genetics. 2020;52(8):754–758.

[25] Do DB Hier, Obafemi-Ajayi T. Balanced Benchmarking of Zero-Shot and RAG Approaches for Biomedical Term Normalization. In: Proceedings of the 2025 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). IEEE; 2025. p. 1–6.

[26] Robinson PN. Deep phenotyping for precision medicine. Human mutation. 2012;33(5):777–780.

[27] Cherry J, Drabkin H, Ebert D, Feuermann M, Gaudet P, Hill D, et al. The gene ontology knowledgebase in 2023. Genetics (Austin). 2023;224(1):iyad031.

[28] UniProt Consortium. UniProt: a hub for protein information. Nucleic acids research. 2015;43(D1):D204–D212.

[29] Eyre TA, Ducluzeau F, Sneddon TP, Povey S, Bruford EA, Lush MJ. The HUGO gene nomenclature database, 2006 updates. Nucleic acids research. 2006;34(suppl_1):D319–D321.

[30] The Jackson Laboratory.: HPO Disease Annotations. Accessed: 2025-01-15. https://hpo.jax.org/data/annotations.

[31] Jackson R, Matentzoglu N, Overton JA, Vita R, Balhoff JP, Buttigieg PL, et al. OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. Database. 2021;2021:baab069.

[32] Gene Ontology Consortium.: Gene Ontology Annotations. Accessed: 2025-01-15. https://current.geneontology.org/annotations/goa_human.gaf.gz.

[33] UniProt Consortium.: UniProt Human Proteome. Accessed: 2025-01-15. https://www.uniprot.org/help/human_proteome.

[34] UniProt Consortium.: UniProt Swiss-Prot Data. Accessed: 2025-01-15. https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.dat.gz.

[35] Wain HM, Bruford EA, Lovering RC, Lush MJ, Wright MW, Povey S. Guidelines for human gene nomenclature. Genomics. 2002;79(4):464–470.

[36] OBO Foundry.: OBO Foundry. Accessed: 2025-01-15. https://obofoundry.org/.

[37] National Center for Biomedical Ontology.: NCBO BioPortal. Accessed: 2025-01-15. https://bioportal.bioontology.org/.

[38] Zipf GK. Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Mansfield, CT: Martino Publishing; 2012. Originally published in 1949 by Addison-Wesley Press, Cambridge, MA.

[39] Shen T, Lee A, Shen C, Lin CJ. The long tail and rare disease research: the impact of next-generation sequencing for rare Mendelian disorders. Genetics research. 2015;97:e15.

[40] Zhang C, Almpanidis G, Fan G, Deng B, Zhang Y, Liu J, et al. A systematic review on long-tailed learning. IEEE Transactions on Neural Networks and Learning Systems. 2025;.

[41] Yi X, Lever J, Bryson K, Meng Z. Can We Edit LLMs for Long-Tail Biomedical Knowledge? arXiv preprint arXiv:250410421. 2025;.

[42] Do T.: A Retrieval Augmented Approach to Improving Accuracy of Biomedical Term Normalization by Large Language Models. Master's thesis, Missouri State University. https://bearworks.missouristate.edu/theses/4045. Available from: https://bearworks.missouristate.edu/theses/4045.

[43] Beck J, Sequeira E. PubMed Central (PMC): An archive for literature from life sciences journals. The NCBI Handbook. 2003;.

[44] Together AI.: LLaMA 3.1 Models. Accessed: 2025-01-15. https://www.together.ai/.

[45] OpenAI.: GPT-4o. Accessed: 2025-01-15. https://openai.com/.

[46] Together AI.: LLaMA 3.1 8B Instruct. Accessed: 2025-01-15. https://www.together.ai/.

[47] Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:210609685. 2021;.

[48] Ni S, Kong X, Li C, Hu X, Xu R, Zhu J, et al. Training on the benchmark is not all you need. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39; 2025. p. 24948–24956.

[49] Bridgman PW. The logic of modern physics. vol. 3. Macmillan; 1927.

[50] Wu Q, Das S, Amani M, Ghosh B, Khan MA, Gummadi KP, et al. Rote Learning Considered Useful: Generalizing over Memorized Data in LLMs. arXiv preprint arXiv:250721914. 2025;.

[51] Chang H, Park J, Ye S, Yang S, Seo Y, Chang DS, et al. How do large language models acquire factual knowledge during pretraining? Advances in neural information processing systems. 2024;37:60626–60668.

[52] Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, et al. Scaling laws for neural language models. arXiv preprint arXiv:200108361. 2020;.

[53] Lu X, Li X, Cheng Q, Ding K, Huang X, Qiu X. Scaling laws for fact memorization of large language models. arXiv preprint arXiv:240615720. 2024;.

[54] Ghosal G, Hashimoto T, Raghunathan A. Understanding finetuning for factual knowledge extraction. arXiv preprint arXiv:240614785. 2024;.

[55] Gekhman Z, David EB, Orgad H, Ofek E, Belinkov Y, Szpektor I, et al. Inside-out: Hidden factual knowledge in LLMs. arXiv preprint arXiv:250315299. 2025;.

[56] Bennane Y. The lexicalization of acronyms in English: The case of third year EFL students, Mentouri University-Constantine. Revue EXPRESSIONS. 2017;.

[57] Periti F, Montanelli S. Lexical semantic change through large language models: a survey. ACM Computing Surveys. 2024;56(11):1–38.

[58] Dai D, Dong L, Hao Y, Sui Z, Chang B, Wei F. Knowledge neurons in pretrained transformers. arXiv preprint arXiv:210408696. 2021;.

[59] Niu J, Liu A, Zhu Z, Penn G. What does the knowledge neuron thesis have to do with knowledge? arXiv preprint arXiv:240502421. 2024;.

Author Information

Author Affiliations

Suswitha Pericharla

Computer Science Department, Missouri State University, Springfield, MO, USA.

Daniel B. Hier
ORCID id: https://orcid.org/0000-0002-6179-0793
Engineering Program, Missouri State University, Springfield, MO, USA.
Department of Neurology and Rehabilitation, University of Illinois, Chicago, IL, USA.

Tayo Obafemi-Ajayi
ORCID id: https://orcid.org/0000-0002-0155-9733
Engineering Program, Missouri State University, Springfield, MO, USA.

Author Contributions

S.P., D.B.H., and T.O. conceived and designed the study. D.B.H. developed the data processing workflow while S.P. performed the experimental analysis. All authors contributed to writing and editing of the manuscript. T.O. supervised the project.

Corresponding Author

Correspondence to Dr. Tayo Obafemi-Ajayi (TayoObafemiAjayi@MissouriState.edu).

Declarations

Ethics Approval and Consent to Participate

This study did not involve human subjects or identifiable human data.

Consent for Publication

All authors have approved the final manuscript for publication.

Availability of Data and Materials

All data analyzed during this study are available on the Computational Learning Systems Lab github page: https://github.com/clslabMSU. Code used in the analysis is also available on the github page.

Competing Interests

The authors declare that they have no competing interests.

Funding

None.

Acknowledgements

None.