Beyond Generation: Unlocking Zero-Shot Entity Representations in Frozen LLMs via Decoupled Probing

Anonymous ACL submission

Abstract

Open-domain zero-shot Named Entity Recognition (NER) relying on generative Large Language Models (LLMs) often suffers from prohibitive autoregressive latency and formatting hallucinations. While internal representation probing offers a generation-free alternative, it faces a severe layer-wise dilemma: early layers capture crisp boundaries but lack semantic depth, whereas deep layers possess rich taxonomic knowledge but blur boundaries due to over-contextualization. To resolve this, we propose **DEEP** (**D**ecoupled **E**ntity **E**xtraction **P**robe), a non-autoregressive framework operating directly on frozen LLM hidden states. DEEP explicitly decouples the extraction pipeline by utilizing early layers for class-agnostic boundary detection and routing the extracted spans to deep layers for prompt-aware semantic typing. Extensive evaluations across diverse zero-shot benchmarks demonstrate that DEEP achieves substantial F_1 improvements (up to +36.2% absolute) while delivering a 13.2$\times$ inference speedup over standard generative baselines.

1 Introduction

Named Entity Recognition (NER) has fundamentally shifted towards open-domain, zero-shot extraction to handle unrestricted entity types across diverse domains. Contemporary approaches predominantly rely on the generative capabilities of Large Language Models (LLMs) via in-context prompting (Wang et al., 2025) or instruction tuning (Wang et al., 2023). However, this generation-based paradigm introduces systemic bottlenecks: vulnerability to instruction deviation, contextual hallucinations (e.g., shifted boundaries or fabricated types) (Huang et al., 2025), and the prohibitive computational latency of autoregressive decoding.

To bypass these generative limitations, recent studies explore LLM internal representations. Interpretability probes, such as ToMMeR (Morand et al., 2026), demonstrate that early transformer layers naturally encode class-agnostic entity boundaries. Yet, shallow probes suffer from exceptionally low precision due to the absence of semantic guidance. Conversely, directly probing deeper layers for zero-shot semantic typing is non-trivial due to the widely recognized phenomenon of **over-contextualization** (Ethayarajh, 2019). Driven by representation anisotropy in deep self-attention, the features of adjacent tokens become highly mixed. Consequently, non-entity or partially overlapping tokens absorb nearby semantic signals, severely blurring structural boundaries. This context diffusion produces hidden states nearly identical to ground-truth entities, preventing traditional probes from achieving precise demarcation.

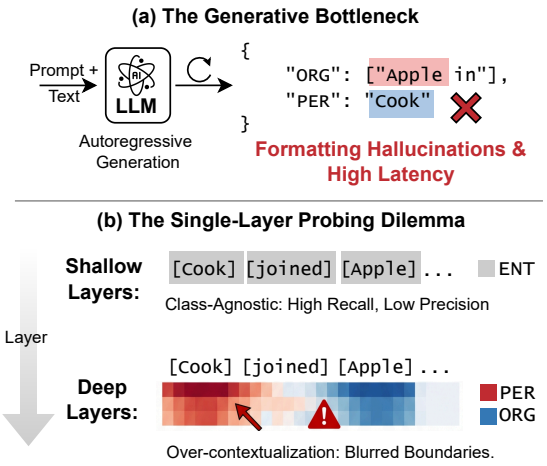

Figure 1: Bottlenecks of zero-shot NER. **(a) Generative Paradigms** suffer from high latency and hallucinations. **(b) Single-Layer Probing Dilemma**: A severe layer-wise trade-off.

To break this precision-recall bottleneck, we propose **DEEP** (**D**ecoupled **E**ntity **E**xtraction **P**robe), a non-autoregressive framework operating directly on frozen LLM hidden states. Rather than struggling against deep-layer over-contextualization,

DEEP cleverly circumvents its negative impact on boundary demarcation by explicitly decoupling the pipeline. We utilize early layers for class-agnostic span detection, capitalizing on their localized, unmixed representations to ensure high boundary recall and sharpness. Subsequently, we leverage deeper layers exclusively for prompt-aware typing. In this typing stage, the dense contextualization of deep layers transforms into an advantage, providing rich semantic cues for precise categorization of the previously demarcated spans.

The primary contributions of this work are summarized as follows:

- **Decoupled Probing Framework:** We propose DEEP, a generation-free paradigm that achieves zero-shot open-domain NER by explicitly decoupling shallow-layer boundary detection and deep-layer semantic typing.
- **Exploiting Layer-wise Dynamics:** We provide a novel perspective on representation probing by utilizing early layers to preserve boundary sharpness against context diffusion, while harnessing deep-layer over-contextualization for rich semantic typing.
- **Superior Efficiency and Accuracy:** Extensive evaluations across diverse zero-shot benchmarks demonstrate that DEEP significantly outperforms state-of-the-art baselines in both extraction precision and inference latency.

2 Related Work

Large Language Models for NER Traditionally, NER relied on discriminative pre-trained models requiring extensive task-specific fine-tuning (Devlin et al., 2019; Lample et al., 2016). The advent of Large Language Models (LLMs) shifted the paradigm toward zero-shot generative extraction via In-Context Learning (Wang et al., 2025) and instruction tuning (Wang et al., 2023). However, autoregressive generation inherently introduces severe computational latency. Furthermore, under zero-shot constraints, generative models remain highly vulnerable to contextual hallucinations, instruction-following degradation, and altered boundaries (Huang et al., 2025; Ma et al., 2023).

Probing and Representation Engineering To bypass generative bottlenecks, Representation Engineering (RepE) (Zou et al., 2025) explores the rich structural signals inherently encoded in LLM hidden states, ranging from noun-phrase boundaries (Clark et al., 2019) to induction mechanisms (Feng and Steinhardt, 2024). Early supervised probes (Popovic and Färber, 2024) successfully localized these states but lacked zero-shot transferability. While generalist extractors like GLiNER (Zaratiana et al., 2024) achieve open-domain transfer, they rely heavily on bi-directional encoders that diverge computationally from the causal mechanism of modern LLMs. Recently, lightweight probes like ToMMeR (Morand et al., 2026) have demonstrated that shallow layers can efficiently localize class-agnostic entity boundaries. Yet, extending this passive extraction to precise zero-shot semantic typing remains a critical open challenge.

Contrastive Learning Contrastive learning aligns text spans and semantic labels in a shared representation space (Huang et al., 2022; Das et al., 2022). To manage the extreme imbalance of non-entities, standard paradigms rely on Hard Example Mining (HEM) (Robinson et al., 2021; Xiong et al., 2020). However, applying HEM to the deep hidden states of frozen LLMs triggers a unique failure mode due to representation anisotropy (Ethayarajh, 2019). Intense contextualization in deep transformers renders adjacent tokens highly homogeneous. In this dense space, forcing models to penalize partially overlapping boundaries (HEM) injects contradictory noise, exacerbating boundary blurring (Zhu and Li, 2022). This fundamental limitation necessitates an architectural paradigm shift, directly motivating our structurally decoupled DEEP pipeline.

3 Methodology

In this section, we present **DEEP** (**D**ecoupled **E**ntity **E**xtraction **P**robe), a non-autoregressive framework that extracts zero-shot entity representations directly from the hidden states of frozen Large Language Models (LLMs). DEEP explicitly decouples the NER pipeline into two orthogonal stages: (1) **Class-Agnostic Span Detection**, which maximizes boundary recall using unmixed shallow features, and (2) **Prompt-Aware Entity Typing**, which leverages deep contextualized representations for precise semantic categorization.

3.1 Stage 1: Class-Agnostic Span Detection

The first stage aims to localize potential entity boundaries with high recall and minimal compu-

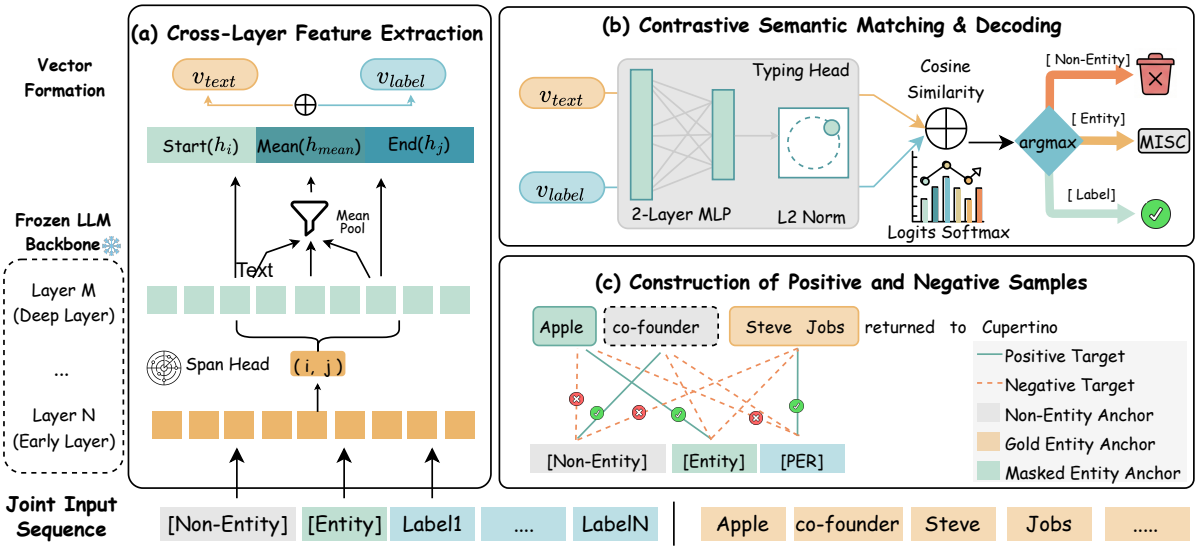

Figure 2: The overall architecture of the DEEP framework. **(a) Cross-Layer Feature Extraction:** Early layers detect class-agnostic boundaries (i, j), which are projected onto deep layers to form dense representations. **(b) Semantic Matching & Decoding:** Features are projected into a shared space and routed to specific labels, an `[Entity]` fallback, or a `[Non-Entity]` discard class. **(c) Contrastive Optimization:** Illustrates the construction of positive and negative span samples as anchors, paired with their corresponding positive and negative label targets.

tational overhead, isolated from complex prompt instructions. We adopt the lightweight token-matching architecture, ToMMeR (Morand et al., 2026), as our off-the-shelf span detector.

Given a raw input sequence, this frozen detector is attached to an early intermediate layer of the LLM backbone. By extracting representations before dense contextualization occurs, we capture highly localized and crisp token boundaries. Instead of generating hard predictions, this stage acts as a high-recall radar, outputting a comprehensive set of candidate spans. Each candidate is strictly anchored to its absolute token coordinates and assigned a confidence score. These extracted boundary coordinates are subsequently passed to Stage 2.

3.2 Stage 2: Prompt-Aware Entity Typing

In Stage 2, the extracted candidate spans are classified by a Typing Head. We project the exact boundary coordinates from Stage 1 onto the deep-layer hidden states of the LLM. Here, the dense contextualization—which previously hindered boundary detection—becomes highly advantageous, providing the rich semantic cues necessary for zero-shot typing.

Joint Prompt Formulation To condition the extraction process on the target categories without introducing multi-pass latency, we concatenate the structured label taxonomy and the raw text tokens into a single unified input sequence:

$$X_{\text{joint}} = \Big[\underbrace{\texttt{[Non-Entity]}, \texttt{[Entity]}, \text{Type}_1, \ldots,}_{\text{Label Taxonomy Prefix}} \underbrace{\text{Type}_N, \text{Token}_1, \ldots, \text{Token}_L}_{\text{Raw Text Sequence}}\Big] \tag{1}$$

By prepending the label taxonomy to the input, the target categories serve as a semantic context. During a single forward pass, the text hidden states naturally incorporate this taxonomy information. This unified formulation allows us to directly extract taxonomy-aware text span representations alongside the label representations for the subsequent classification.

Span Representation Fusion Using the absolute coordinates from Stage 1, we map each candidate to the tokenized, prompt-prepended sequence. For a mapped text span or label span ranging from token index i to j, let $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K]$ denote the deep hidden states of the sequence. The fused representation $\mathbf{v}_{span} \in \mathbb{R}^{3d}$ is constructed via:

$$\mathbf{v}_{span} = [\mathbf{h}_i \oplus \mathbf{h}_{mean} \oplus \mathbf{h}_j] \tag{2}$$

where $\mathbf{h}_{mean}$ denotes the mean-pooled representation of the hidden states within the span, while

$\mathbf{h}_i$ and $\mathbf{h}_j$ explicitly capture the boundary signals. We apply this identical fusion strategy to the label spans within the prefix to derive the corresponding label representations, denoted as $\mathbf{v}_{label}$.

Contrastive Semantic Matching To measure the semantic compatibility between the m-th candidate span and the c-th target label, we project their representations into a shared semantic space via a shared multi-layer perceptron (MLP). Specifically, $\mathbf{v}_{span_m}$ and $\mathbf{v}_{label_c}$ are processed by the same 2-layer MLP and subsequently ℓ_2-normalized. The final compatibility score $z_{m,c}$ is computed via a temperature-scaled cosine similarity:

$$z_{m,c} = \tau \cdot \left(\frac{\text{MLP}(\mathbf{v}_{span_m})}{\|\text{MLP}(\mathbf{v}_{span_m})\|_2} \cdot \frac{\text{MLP}(\mathbf{v}_{label_c})}{\|\text{MLP}(\mathbf{v}_{label_c})\|_2} \right)$$

where τ is a learnable scalar that dynamically calibrates the sharpness of the similarity distribution. By computing this metric across all prompt labels, we obtain a comprehensive score vector $\mathbf{z}_m \in \mathbb{R}^{|\mathcal{C}|}$ for the m-th candidate span, enabling precise multi-class zero-shot typing.

3.3 Minimalist Contrastive Optimization

To map the extracted span and label representations into a discriminative semantic space without relying on heavy parameter updates, we employ a minimalist optimization strategy. Because the underlying LLM remains frozen, our objective focuses purely on fine-tuning the projection MLPs via contrastive alignment, explicitly decoupling the calibration of valid entities from non-entities.

Dual-Tier Contrastive Objective Our typing head functions as a multi-class classifier over the expanded prompt taxonomy, where the zeroth index is strictly reserved for the generic `[Non-Entity]` label, and the first index corresponds to the generic `[Entity]` placeholder. To foster robust open-domain generalization and seamlessly accommodate unseen or miscellaneous categories during inference, we introduce a **Label Dropping** mechanism during training. Specifically, we randomly discard a fixed proportion of specific target labels from the taxonomy prefix. For optimization, we construct two distinct sample sets:

- **Gold Spans ($\mathcal{B}_+$):** For a ground-truth entity m, its positive target c_m depends on the label dropping outcome. If its corresponding specific semantic label remains available in the prefix, c_m points to that specific category. Conversely, if its specific label has been randomly discarded, its positive target c_m defaults to the generic `[Entity]` label ($c = 1$). All other remaining labels within the prompt taxonomy serve as in-batch negatives.
- **Sampled Negative Spans ($\mathcal{B}_-$):** We adopt a stochastic negative sampling strategy by randomly selecting non-entity boundaries from the input sequence. Empirically, we find that this simple random sampling yields more robust representations than heuristic hard-negative mining (a phenomenon we analyze in detail in the experiments section). For these randomly sampled spans, the sole positive target is the `[Non-Entity]` label.

Applying a balanced cross-entropy formulation, we compute the contrastive loss for both subsets:

$$\mathcal{L}_{\text{pos}} = -\frac{1}{|\mathcal{B}_+|} \sum_{m \in \mathcal{B}_+} \log \left(\frac{\exp(z_{m,c_m})}{\sum_{k \in \mathcal{C}} \exp(z_{m,k})} \right)$$

$$\mathcal{L}_{\text{neg}} = -\frac{1}{|\mathcal{B}_-|} \sum_{m \in \mathcal{B}_-} \log \left(\frac{\exp(z_{m,0})}{\sum_{k \in \mathcal{C}} \exp(z_{m,k})} \right)$$

By optimizing their mean $\mathcal{L} = \frac{1}{2}(\mathcal{L}_{\text{pos}} + \mathcal{L}_{\text{neg}})$, we ensure that accurate semantic typing (including open-domain generalization) and robust boundary rejection are weighted equally during parameter updates.

3.4 Inference Strategy: Decoupled Decoding and Semantic Fallback

During zero-shot inference, DEEP elegantly divides the decision-making process into two orthogonal mechanisms.

Threshold-Argmax Decoupling First, Stage 1 proposes a set of candidate boundaries by applying a heuristic decoding strategy with a flexible confidence threshold, acting as a high-recall filter. Subsequently, Stage 2 performs deterministic categorization over these candidates without requiring additional tuning. We directly compute the multi-class argmax over the logit distribution: $\hat{c}_m = \arg\max_{c \in \mathcal{C}} z_{m,c}$. If $\hat{c}_m = 0$ (i.e., `[Non-Entity]`), the candidate is safely discarded; otherwise, it is formally extracted as a valid entity of type $\hat{c}_m$.

Semantic Fallback Mapping Certain dataset-specific classes, such as the `MISC` category in

CoNLL2003, lack a distinct semantic profile. Because our generic `[Entity]` token naturally captures valid entities that do not align with concrete types (e.g., `PER`, `ORG`), it serves as an ideal semantic equivalent. During post-processing, any candidate predicted as `[Entity]` is deterministically mapped to the `MISC` category, offering a simple and robust semantic fallback mechanism.

4 Experiments

4.1 Experimental Setup

Implementation Details To guarantee a strictly fair comparison, all internal probing frameworks and generative baselines are anchored on the frozen **Llama-3.2-1B** (AI@Meta, 2024) backbone. For our DEEP framework, Stage 1 explicitly extracts span boundaries from Layer 5, while Stage 2 conducts prompt-aware semantic typing at Layer 10. The Stage 2 typing head is optimized on Pile-NER via AdamW. Due to space constraints, exhaustive implementation details—including hyperparameter configurations, negative sampling ratios, and hardware specifications—are systematically detailed in Appendix F.

Evaluation Benchmarks We evaluate zero-shot transferability across diverse linguistic distributions using three downstream benchmarks: **CoNLL2003** (Tjong Kim Sang and De Meulder, 2003) (general news domain), **CrossNER** (Liu et al., 2020) (comprising five specialized out-of-domain sets), and **GENIA** (Kim et al., 2003) (highly technical biomedical literature). To optimize main-text efficiency, exhaustive dataset profiles, entity taxonomies, and corpus statistics are systematically detailed in Appendix G.

Baseline Paradigms We evaluate DEEP against two established paradigms. For **generative baselines**, we compare against *Few-Shot ICL* (via 5-shot JSON prompting) and a *Fine-Tuned* model explicitly instruction-tuned on the Pile-NER corpus (Wang et al., 2023). For the **probing baseline**, we employ *ToMMeR* (Morand et al., 2026), a state-of-the-art interpretive probe that extracts class-agnostic entity boundaries exclusively from shallow transformer layers.

4.2 Main Quantitative Results

Table 1 presents the definitive zero-shot extraction performance across the general domain (CoNLL2003) and five highly specialized out-of-domain scenarios (CrossNER). The results demonstrate that DEEP establishes a new state-of-the-art for internal representation probing.

Bypassing the Generative Bottleneck The empirical results expose severe systematic vulnerabilities in autoregressive generation paradigms under zero-shot and cross-domain constraints. The Few-Shot ICL baseline collapses almost entirely across all domains (F1 scores ranging from 0.142 to 0.256), limited by instruction-following degradation and format compliance failures.

A critical observation emerges when comparing the Fine-Tuned generative baseline with our DEEP framework, as both models are trained on the exact same diverse Pile-NER corpus. While explicit instruction tuning stabilizes structured output on the general domain (CoNLL2003 F1 = 0.474), it exhibits severe catastrophic degradation when evaluated against specialized, unseen terminology (e.g., dropping to an F1 of 0.333 on Science). This divergence demonstrates that exposure to massive open-domain data alone is insufficient; generative LLMs tend to overfit surface-level textual distributions. In stark contrast, DEEP acts as an optimal structural conduit, successfully unlocking the rich, open-domain taxonomic knowledge embedded during Pile-NER training without incurring the generative alignment penalty.

Bridging the Precision Gap in Probing The shallow-only probing baseline, ToMMeR, yields an exceptionally high structural boundary recall (0.966 on CoNLL2003). However, because early layers lack high-level taxonomic semantics, it indiscriminately extracts non-entities, resulting in suboptimal precision (0.281).

DEEP elegantly resolves this precision-recall bottleneck. By utilizing early layers purely as a class-agnostic span radar and offloading categorization to deep-layer prompt-aware hidden states, DEEP effectively filters out false boundaries while maintaining the integrity of the pre-trained semantic manifold. This minimalist architectural decoupling drives precision up to 0.690 (+40.9% absolute increase over ToMMeR) and establishes a commanding F1 score of 0.743 on CoNLL2003.

Exceptional Cross-Domain Generalization A defining strength of the DEEP framework is its robust stability across diverse out-of-domain (OOD) scenarios. Unlike generative models that strug-

Method	CoNLL2003 (General Domain)			CrossNER Zero-Shot (F_1 Score)				
	Precision	Recall	F_1	AI	Literature	Politics	Music	Science
Generative Paradigms (LLM Autoregressive Generation)								
Few-Shot ICL	0.170	0.345	0.228	0.157	0.142	0.256	0.228	0.204
Fine-Tuned (Pile-NER)	0.361	0.687	0.474	0.360	0.380	0.455	0.448	0.333
Probing Paradigms (Internal Representations)								
ToMMeR (Shallow-only)	0.281	**0.966**	0.436	0.489	0.520	0.458	0.548	0.536
DEEP (Ours)	**0.690**	0.805	**0.743**	**0.692**	**0.699**	**0.722**	**0.736**	**0.690**

Table 1: Main zero-shot and out-of-domain NER extraction performance comparison using Llama-3.2-1B. DEEP achieves a superior trade-off by breaking the low-precision boundary bottleneck of representation probes while fully bypassing generative format hallucinations.

gle heavily with highly specialized vocabularies, DEEP maintains a tightly bounded, exceptionally high F1 score between 0.692 and 0.736 across all five CrossNER domains. Most notably, DEEP achieves a staggering +35.7% absolute F1 improvement over the Fine-Tuned baseline on the Science domain. This robust zero-shot generalization validates that our decoupled internal probing extracts fundamental semantic concepts of entity tracking, rather than memorizing domain-specific text patterns.

4.3 Ablation Study

To isolate and quantify the contribution of specific mechanisms in our decoupled framework, we conduct an ablation study on the CoNLL2003 and CrossNER (AI) datasets (Table 2).

Architecture Variant	CoNLL2003		CrossNER (AI)	
	Precision	F_1	Precision	F_1
Full DEEP Framework	**0.690**	**0.743**	**0.649**	**0.692**
Semantic Mapping Strategy				
w/o Semantic Fallback	0.595	0.668	0.575	0.634
Span Representation (Vector Formation)				
w/o Mean Pooling (Boundary-only)	0.641	0.722	0.516	0.587
w/o Boundary Tokens (Mean-only)	0.690	0.731	0.623	0.671

Table 2: Ablation study of DEEP components. Both the semantic fallback mechanism and the comprehensive span feature formation are critical for maintaining high precision in cross-domain scenarios.

Effect of Semantic Fallback We evaluate the necessity of our semantic mapping strategy (*w/o* Semantic Fallback) by removing the generic `[Entity]` token from the prompt taxonomy. Instead, we force the model to directly predict the dataset-specific `MISC` label. As shown in Table 2, this modification induces a substantial performance degradation, dropping the F_1 score by 7.5% absolute on CoNLL2003 and a severe 5.8% on CrossNER (AI). This validates our hypothesis: forcing an LLM to align with vaguely defined, dataset-specific jargon disrupts its zero-shot semantic mapping capabilities. Conversely, employing a natural ontological proxy (i.e., `[Entity]`) effectively anchors the representations, preserving high typing accuracy even in specialized, out-of-domain scenarios.

Impact of Span Representation We further ablate the cross-layer vector formation strategy. Relying solely on boundary tokens (*w/o* Mean Pooling) causes a severe performance collapse, most notably plunging the CrossNER (AI) F_1 score by a massive 10.5% absolute (from 0.692 to 0.587). This indicates that in deep transformer layers, boundary coordinates alone lack the requisite semantic density to classify complex, out-of-domain entities. Furthermore, relying solely on the internal contextual span (*w/o* Boundary Tokens) also yields consistent performance degradation. This empirically demonstrates that concatenating explicit boundary coordinates with the internal semantic aggregate is essential to counteract boundary blurring and preserve feature sharpness.

4.4 Scaling on Complex Domains and World Knowledge Elicitation

Since general-domain benchmarks saturate early, we evaluate DEEP on highly specialized datasets using **Phi-2 (2.7B)** (Mo et al., 2024) and **Llama-3.2-3B** (AI@Meta, 2024) to rigorously test deep-layer world knowledge elicitation. Table 3 presents three representative complex domains: **GENIA** (Bio-medicine), **Science**, and **Politics** (full results across all six domains are in Appendix B).

Eliciting Deep Taxonomic Knowledge in Specialized Domains The data in Table 3 reveals a pro-

Domain	Llama-3.2-3B					Phi-2 (2.7B)				
	DEEP (Ours)			ToMMeR		DEEP (Ours)			ToMMeR	
	Typing Acc.	F_1	Δ	Recall	F_1	Typing Acc.	F_1	Δ	Recall	F_1
GENIA (Bio)	**0.8447**	**0.5477**	**+0.362**	0.8816	0.1853	**0.8437**	**0.5093**	**+0.238**	0.8955	0.2706
Science	0.8268	0.7246	+0.215	0.9950	0.5095	0.8243	0.6911	+0.199	0.9310	0.4921
Politics	0.8461	0.7301	+0.265	0.9622	0.4649	0.8630	0.7335	+0.291	0.9413	0.4417

Table 3: Zero-shot performance on representative complex domains. Δ denotes the absolute F_1 improvement over the shallow-only ToMMeR baseline. Larger models exhibit profound deep-layer taxonomic knowledge in specialized domains, particularly evident in the biomedical GENIA dataset.

found insight into the mechanics of representation probing. On highly specialized datasets like the biomedical **GENIA** benchmark, the shallow-only baseline (ToMMeR) catastrophically fails (e.g., $F_1 = 0.1853$ on Llama-3B), despite maintaining a relatively high recall (0.8816). This indicates that while early layers can detect structural noun phrases, they possess zero understanding of complex biomedical jargon.

However, by routing these agnostic spans into DEEP's Stage 2 deep-layer typing head, the performance experiences an explosive surge. On Llama-3.2-3B, DEEP boosts the GENIA F_1 score by an astonishing **+36.2%** absolute points, driven by a deep-layer typing accuracy of **84.47%**.

Parameter Scaling Unlocks Domain Generalization This breakthrough empirically proves that large-scale pre-training endows LLMs with profound world knowledge across niche domains (such as Bio-medicine and Science). Shallow probing inherently bottlenecks this capacity. Our decoupled DEEP framework acts as an optimal knowledge elicitation engine: as the backbone model scales from 1B to 3B, its deep-layer semantic manifold becomes increasingly enriched, naturally translating into state-of-the-art zero-shot categorization on the most challenging cross-domain NER tasks.

4.5 Layer Dynamics and Internal Representation Evolution

A crucial inquiry in representation engineering is how semantic typing capacity evolves across the LLM's internal depth. To investigate this, we sweep the extraction layer for our Stage 2 typing head across the Llama-3.2-1B backbone (Layers 2 to 14) while keeping the Stage 1 span detection fixed, as illustrated in Figure 3.

From Eager Recall to Cautious Precision The layer-wise trajectory reveals a fascinating evolution of internal representations. Shallow layers (e.g., Layer 4) exhibit highly "eager" behavior: they achieve peak entity recognition accuracy (88.2%) and overall recall (0.852), but struggle with the lowest non-entity rejection rate (74.1%). This indicates strong structural sensitivity but immature semantic discrimination. Conversely, as we probe deeper (e.g., Layer 10), representations become notably "cautious." The model achieves its peak precision (0.690) and non-entity rejection accuracy (79.4%), successfully filtering out high-confidence false positives passed from Stage 1.

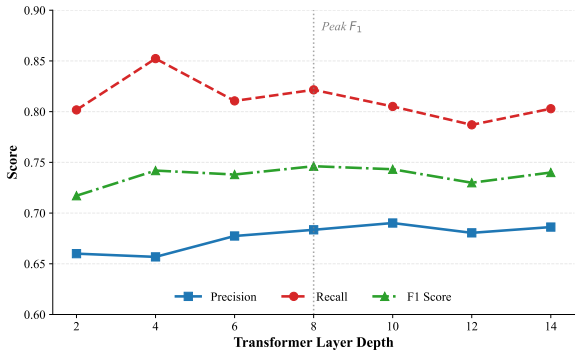

Figure 3: Performance evolution across intermediate layers. Shallow layers exhibit an "eager" high-recall behavior, whereas middle-to-late layers develop "cautious" semantic precision. The optimal precision-recall trade-off naturally emerges around Layer 8.

The Golden Middle and Architectural Robustness The optimal trade-off between structural recall and semantic precision materializes in the middle-to-late layers (Layers 8-10, peaking at an F_1 of 0.746). This corroborates the hypothesis that deep semantic taxonomies fully crystallize before the absolute final layers begin projecting back to the vocabulary space. Notably, the overall F_1 performance remains remarkably stable (≥ 0.717) across all tested depths. This architectural robustness confirms that DEEP's effectiveness is rooted in its explicit decoupling mechanism, rather than a brittle reliance on a specific magic layer.

4.6 The Pitfalls of Forced Regularization in Deep Representations

While advanced negative sampling techniques like Online Hard Example Mining (OHEM) and Boundary-Aware Negative Mining (BANM) are standard in supervised NER, we investigate whether they translate effectively to frozen LLMs. We compare our **Minimalist** objective (pairing gold entities with randomly sampled non-entities) against two aggressive variants: **+OHEM** (penalizing Stage 1's high-confidence false positives) and **+BANM** (penalizing structurally perturbed boundaries). Table 4 presents the results on CoNLL2003 and Science (full results are in Appendix C).

Sampling Strategy	CoNLL2003			CrossNER (Science)		
	Precision	Recall	F_1	Precision	Recall	F_1
Minimalist (Ours)	**0.690**	0.805	**0.743**	**0.648**	0.737	**0.690**
+ OHEM	0.632	0.803	0.707	0.568	0.736	0.641
+ BANM	0.609	**0.862**	0.714	0.528	**0.819**	0.642

Table 4: Impact of negative sampling strategies. Aggressive regularization (OHEM, BANM) severely degrades precision, validating our minimalist approach.

Preserving the Pre-trained Semantic Manifold Counter-intuitively, OHEM and BANM degrade performance, notably plummeting precision on the Science domain (from 0.648 to 0.528). This highlights a fundamental difference between training from scratch and probing frozen LLMs. For OHEM, Stage 1's "hard negatives" are often structurally valid open-domain entities that simply lack annotations in the target ontology; penalizing them injects contradictory noise that overrides the LLM's pre-trained semantic priors. Similarly, BANM forces deep layers to perform microscopic boundary discrimination—a task they are inherently ill-suited for due to context diffusion. These findings validate our minimalist philosophy: avoiding forced regularization preserves the deep layers' pristine classification space, optimally unleashing their zero-shot capabilities.

4.7 Inference Efficiency and Latency

A core motivation for probing internal representations is circumventing the severe latency bottleneck of autoregressive generation. Table 5 profiles the end-to-end inference latency using a standard batch size of 1.

In a standard PyTorch environment, generative paradigms suffer from severe decoding delays (e.g., 589.62 ms for the FT baseline). By completely eliminating token-by-token generation, DEEP requires only a single forward pass, slashing the total latency to just **44.42 ms** (15.94 ms for Stage 1; 28.48 ms for Stage 2). This delivers a staggering **13.2x speedup** over vanilla FT.

Method	Inference Engine	Latency (ms)	Speedup
Few-Shot ICL	Vanilla PyTorch	1107.40	0.5×
	vLLM	109.01	5.4×
Fine-Tuned (FT)	Vanilla PyTorch	589.62	1.0× (Ref)
	vLLM	29.21	20.1×
DEEP (Ours)	Vanilla PyTorch	**44.42**	**13.2×**

Table 5: End-to-end inference latency. By eliminating autoregressive decoding, vanilla DEEP achieves a 13.2× speedup over the vanilla FT baseline, remaining highly competitive even without specialized serving frameworks like vLLM.

Furthermore, while generative models inherently rely on memory-intensive serving engines (e.g., vLLM with PagedAttention (Kwon et al., 2023)) to achieve practical deployment speeds (29.21 ms), our out-of-the-box DEEP implementation remains highly competitive. This demonstrates that DEEP resolves the latency bottleneck fundamentally at the algorithmic level, bypassing the need for complex, heavily engineered infrastructure.

5 Conclusion

In this paper, we proposed **DEEP**, a Decoupled Entity Extraction Probe that resolves the prohibitive latency of autoregressive generation and the precision bottlenecks of standard representation probing. By explicitly decoupling the NER pipeline and leveraging shallow layers for class-agnostic span detection and deep layers for prompt-aware semantic typing, DEEP effectively bypasses the boundary-blurring effects of deep layer over-contextualization.

Extensive evaluations demonstrate that DEEP establishes a new zero-shot knowledge elicitation, achieving massive F_1 improvements while delivering a 13.2× latency speedup over generative baselines. Furthermore, our analysis reveals that aggressive negative sampling actively corrupts the LLM's pre-trained semantic manifold, validating our optimization strategy. Ultimately, DEEP demonstrates that non-generative, decoupled probing is a highly efficient and scalable trajectory, proving that the most direct path to unlocking internal LLM knowledge often requires the least interference.

6 Limitations

While the DEEP framework establishes a highly efficient zero-shot NER paradigm, we acknowledge several limitations that provide avenues for future research. First, our layer-selection heuristic (e.g., $\lfloor 0.6 \times L \rfloor$) lacks formal mechanistic interpretability, warranting deeper investigation into exactly how internal semantic manifolds crystallize. Second, due to computational constraints, our evaluations are anchored on models up to the 3B scale; whether the observed boundary diffusion phenomena scale consistently to massive LLMs (e.g., 70B+) remains an open question. Furthermore, as an explicitly decoupled pipeline, DEEP is inherently susceptible to cascading errors—any boundaries missed by the Stage 1 span radar cannot be recovered by the Stage 2 typing head. Finally, our current boundary detection mechanism primarily targets flat entities, requiring future structural extensions to effectively accommodate complex nested or discontinuous entity structures.

References

AI@Meta. 2024. Llama 3 model card.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Sarkar Snigdha Sarathi Das, Arzoo Katiyar, Rebecca Passonneau, and Rui Zhang. 2022. CONTaiNER: Few-shot named entity recognition via contrastive learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6338–6353, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Jiahai Feng and Jacob Steinhardt. 2024. How do language models bind entities in context? *Preprint*, arXiv:2310.17191.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Yucheng Huang, Kai He, Yige Wang, Xianli Zhang, Tieliang Gong, Rui Mao, and Chen Li. 2022. COPNER: Contrastive learning with prompt guiding for few-shot named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2515–2527, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(Suppl 1):i180–i182.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Preprint*, arXiv:2309.06180.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2020. Crossner: Evaluating cross-domain named entity recognition. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI-21)*, volume 35. AAAI Press.

Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. 2023. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10572–10601, Singapore. Association for Computational Linguistics.

Marah Mo, Songtao Yang, Chunting Hsieh, Tsung-Yi Lee, Chaitanya Joshi, Kurt Keutzer, Amir Ghorbani, Mostofa Shoeybi, Mohammad Patwary, Shrimai Narayanan, and 1 others. 2024. Phi-2: The surprising power of small language models. *Preprint*, arXiv:2402.11664. Microsoft Technical Report.

Victor Morand, Nadi Tomeh, Josiane Mothe, and Benjamin Piwowarski. 2026. Tommer – efficient entity mention detection from large language models. *Preprint*, arXiv:2510.19410.

Nicholas Popovic and Michael Färber. 2024. Embedded named entity recognition using probing classifiers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17830–17850, Miami, Florida, USA. Association for Computational Linguistics.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. *Preprint*, arXiv:2010.04592.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. GPT-NER: Named entity recognition via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275, Albuquerque, New Mexico. Association for Computational Linguistics.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, Jihua Kang, Jingsheng Yang, Siyuan Li, and Chunsai Du. 2023. Instructuie: Multi-task instruction tuning for unified information extraction. *Preprint*, arXiv:2304.08085.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *Preprint*, arXiv:2007.00808.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

Enwei Zhu and Jinpeng Li. 2022. Boundary smoothing for named entity recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7096–7108, Dublin, Ireland. Association for Computational Linguistics.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2025. Representation engineering: A top-down approach to ai transparency. *Preprint*, arXiv:2310.01405.

A Reproducibility and Code Availability

In accordance with the double-blind review policy, the repository has been thoroughly anonymized and is available at: https://anonymous.4open.science/r/DEEP-88E5/README2.md.

B Comprehensive Zero-Shot Results on Complex Domains

In Section 4.4, we presented the scaling behavior of DEEP on three representative complex domains. For completeness, Table 6 provides the comprehensive evaluation results across all six specialized out-of-domain datasets (GENIA, AI, Literature, Politics, Music, and Science) using both Llama-3.2-3B and Phi-2.

Consistently across all six domains, DEEP yields massive absolute improvements (Δ) over the shallow-layer baseline by successfully eliciting the deep-layer semantic taxonomies of the backbone models.

C Comprehensive Ablation on Negative Sampling

In Section 4.6, we discussed the detrimental effects of aggressive negative sampling strategies (OHEM and BANM) on the deep-layer semantic manifold, highlighting their impact on CoNLL2003 and the Science domain.

For completeness, Table 6 provides the expanded ablation results across all available CrossNER domains. Across almost all datasets, the injection of OHEM strictly degrades both Precision and Overall Accuracy, further substantiating our hypothesis that forcing a frozen LLM to unlearn high-confidence predictions disrupts its pre-trained world knowledge. (Note: BANM evaluation on certain domains was omitted due to its consistently detrimental impact on precision observed during early trials).

D Appendix: Prompt Used for the Fine-tune Model

```
You are a precise Named Entity Recognition (NER)
      system.

Task:
Extract all named entities that appear verbatim
     in the input text
and assign each one a type from the allowed
     Entity Types.

Text: {text}
Entity Types: {type_list}
Answer:
```

Domain	Llama-3.2-3B					Phi-2 (2.7B)				
	DEEP (Ours)			ToMMeR		DEEP (Ours)			ToMMeR	
	Typing Acc.	F_1	Δ	Recall	F_1	Typing Acc.	F_1	Δ	Recall	F_1
GENIA (Bio)	**0.8447**	**0.5477**	**+0.362**	0.8816	0.1853	**0.8437**	**0.5093**	**+0.238**	0.8955	0.2706
AI	0.8215	0.6913	+0.219	0.9585	0.4720	0.8338	0.6810	+0.241	0.9000	0.4400
Literature	0.8383	0.7061	+0.206	0.9330	0.4995	0.8397	0.6690	+0.198	0.8932	0.4708
Politics	0.8461	0.7301	+0.265	0.9622	0.4649	0.8630	0.7335	+0.291	0.9413	0.4417
Music	0.8287	0.7395	+0.201	0.9325	0.5385	0.8256	0.7229	+0.204	0.9259	0.5189
Science	0.8268	0.7246	+0.215	0.9950	0.5095	0.8243	0.6911	+0.199	0.9310	0.4921

Table 6: Comprehensive zero-shot performance across all tested complex domains. DEEP consistently translates deep-layer world knowledge into high-precision categorization.

Domain	Minimalist DEEP (Ours)				DEEP (+ OHEM)				DEEP (+ BANM)			
	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc
CoNLL2003	0.6902	0.8051	0.7432	0.8058	0.6321	0.8029	0.7073	0.7630	0.6094	0.8623	0.7141	0.7417
Science	0.6481	0.7375	0.6899	0.7833	0.5686	0.7362	0.6416	0.7211	0.5094	0.8488	0.6367	0.6714
AI	0.6499	0.7407	0.6230	0.8000	0.5494	0.7253	0.6252	0.7350	0.4800	0.8270	0.6109	0.6867
Literature	0.6813	0.7171	0.6988	0.8142	0.5963	0.6778	0.6334	0.7531	0.5540	0.7970	0.6536	0.7354
Politics	0.6639	0.7902	0.7216	0.8386	0.5627	0.7850	0.6555	0.7702	0.4735	0.8757	0.6147	0.6894
Music	0.6933	0.7848	0.7362	0.8104	0.6171	0.7509	0.6775	0.7479	0.5695	0.8624	0.6860	0.7240

E Appendix: Prompt Template Used for the NER Model

```
You are a precise Named Entity Recognition (NER)
      system.

Task:
Extract all named entities that appear verbatim
     in the input text
and assign each one a type from the allowed
     schema.

Allowed entity types:
{entity_types}

Rules:
1. Only output entity mentions that are exact
     substrings of the input text.
2. Do not output any entity mention that appears
      only in the few-shot examples.
3. Preserve the original surface form from the
     input text, including casing and punctuation
     .
4. Output each mention once. If the same surface
      form appears multiple times, keep one entry
     .
5. If no valid entities are present, output an
     empty JSON object: {}.
6. Output exactly one JSON object and nothing
     else.

[Few-shot Examples]
{examples}

[Input Text]
Text: "{input_text}"

[Output]
```

F Hyperparameters and Implementation Details

In this appendix, we provide the exhaustive implementation details and hyperparameter configurations required to reproduce the DEEP framework.

Layer Selection Heuristic For Stage 2 (Prompt-Aware Entity Typing), we empirically select an intermediate layer at a depth ratio of approximately 0.6 relative to the total network layers (L). This mid-to-late layer preference effectively avoids both the semantic shallowness of early layers and the task-specific bias (i.e., next-token prediction) of the absolute final layers. Following this $\lfloor 0.6 \times L \rfloor$ heuristic, we anchor the typing head at Layer 10 for Llama-3.2-1B ($L = 16$) and Layer 18 for Llama-3.2-3B ($L = 28$).

Training Setup During optimization, the underlying LLM backbone remains strictly frozen, and we only update the lightweight projection MLPs. All experiments are conducted on NVIDIA A100 GPUs. The detailed hyperparameter configurations across different model scales are summarized in Table 7.

G Dataset Details and Specifications

In this appendix, we present the comprehensive profiles and taxonomies of the evaluation benchmarks utilized in our primary experiments.

Table 7: Experimental Hyperparameters for DEEP (code-verified).

Category	Item	Value
1) Architecture Settings		
	Backbone Models	Llama-3.2-1B
	Stage-1 Layer Index (Span Head)	1B: 5, 3B: 1
	Stage-2 Layer Index (Typing Head)	1B: 10, 3B: 18
	Max Sequence Length	Span stage: 1024; Typing stage: 1024
	Typing Head Projection Dim	256
	Typing Head Trainable Params	1B: 26,219,009; 3B: 58,202,625
	Hidden Size h (from model configs)	1B: 2048, 3B: 3072; $p = 256$
2) Training & Optimization		
	Optimizer	AdamW
	Learning Rate	1×10^{-3}
	Batch Size	4
	Training Epochs	5
	Warmup	300 steps (linear warmup)
	LR Decay	Cosine decay to min_lr_ratio $= 0.1$
	Gradient Clipping	1.0
	Gradient Accumulation Steps	1
	Weight Decay	0.01 (AdamW default)
3) Method-Specific Hyperparameters		
	Negative Sample Size	10
	Hard Negative Strategy	random
	Label Dropping Ratio	0.2 (mask_prob)
	Stage-1 Confidence Threshold	0.5
	Span Decoding Strategy	threshold
	Temperature τ	Learnable; initialized as 0.07
4) Hardware		
	GPU Type	NVIDIA RTX 4090

Layer	Precision	Recall	F_1	Overall Acc	Ent. Acc	Non-Ent. Acc
2	0.660	0.801	0.717	0.787	0.829	0.771
4	0.656	**0.852**	0.741	0.781	**0.882**	0.741
6	0.677	0.810	0.738	0.799	0.838	0.783
8	0.683	0.821	**0.746**	0.790	0.850	0.776
10	**0.690**	0.805	0.743	**0.805**	0.833	**0.794**
12	0.680	0.787	0.729	0.792	0.814	0.783
14	0.686	0.802	0.740	0.787	0.831	0.770

Table 8: Performance evolution across intermediate layers. Shallow layers exhibit an "eager" high-recall behavior, whereas middle-to-late layers develop "cautious" semantic precision, peaking around Layers 8-10.

General and Cross-Domain Benchmarks The **CoNLL2003** dataset serves as our general news baseline, annotated with four standard entity categories: Person (`PER`), Location (`LOC`), Organization (`ORG`), and Miscellaneous (`MISC`). The **CrossNER** benchmark evaluates zero-shot cross-domain adaptation across five highly distinct, non-overlapping domains: Artificial Intelligence (*AI*), Literature, Politics, Music, and Science, each possessing a unique vocabulary distribution and specialized category taxonomy.

GENIA (Complex Biomedical Domain) The **GENIA** corpus (Kim et al., 2003) is a highly specialized, niche benchmark derived from molecular biology literature, specifically compiled from MEDLINE abstracts. It represents an exceptional challenge for zero-shot open-domain NER due to its dense concentration of domain-specific terminology, intricate biochemical nomenclatures, and advanced scientific jargon. The dataset is rigorously annotated with five primary biological entity categories: `DNA`, `RNA`, `Protein`, `Cell_Line`, and `Cell_Type`.

Because extracting these entities requires more than surface-level syntactic pattern matching, GE-

NIA serves as an ideal stress-test for world-knowledge elicitation. While early transformer layers can identify the boundaries of these complex technical phrases (high recall), they completely lack the high-level semantic maturity to categorize them. We leverage the standard test split of GENIA to empirically demonstrate how DEEP effectively unlocks the deep-layer semantic manifold of frozen LLMs to perform precise multi-class zero-shot categorization on highly specialized, out-of-domain scientific data.