Distribution, Dosage, Density: What Dominates a BERT-like Model's Learning to Disambiguate Word Senses?

Anonymous ACL submission

Abstract

Word Sense Disambiguation (WSD) is a critical Natural Language Processing (NLP) task that aims to select the correct meaning of a word based on its context. Recent approaches employ BERT-like models (e.g., BERT, RoBERTa) as their backbone, achieving strong performance on WSD. However, the disambiguation ability of Pretrained Language Models (PLMs) has not shown substantial improvement. Moreover, studies have shown that PLMs already possess some innate WSD ability. A natural question, then, is what factors primarily dominate the further refinement of this ability during fine-tuning. To investigate this issue, in this work, we conduct a probing study to examine the influence of three key factors: distribution, dosage, and density. We find that although distribution affects certain senses, the absolute number of training instances per sense is an equally crucial determinant of performance. Larger dosage reduces the number of intermediate layers needed for contextual understanding and increases disambiguation confidence; on high-dosage datasets, models with fewer layers achieve better performance while reducing training time by 7.97%. We further show that low density of semantic keywords reduces attention to key cues and limits disambiguation accuracy. Replacing the context of 150 instances with annotated keywords improves performance (micro-F1 from 0 to 0.34; macro-F1 from 0 to 0.16), indicating that increasing density helps models better attend to informative cues and enhances WSD capability. Our code is available at `https://anonymous.4open.science/r/BERT-wsd-probing-1686/`.

1 Introduction

Polysemy is a ubiquitous linguistic phenomenon in Natural Language Processing (NLP) that poses a challenge to language models for semantic understanding. Word Sense Disambiguation (WSD), a fundamental NLP task, addresses this by determining the correct meaning of a polysemous word in context (Navigli, 2009). It is commonly formulated as a multi-label classification problem (Raganato et al., 2017), requiring the model to choose a sense from a predefined inventory (typically WordNet (Miller, 1994)) based on the contextual clues, in order to better semantic comprehension and enhance performance on downstream applications like machine translation (Reheman et al., 2025), text summarization (Wei et al., 2025), and information extraction (Chen et al., 2025).

Most WSD approaches rely on high-quality, human-annotated datasets as training corpus, which typically adopt BERT-like models (e.g., BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019)) as their backbone, leveraging their strong capacity for fine-grained semantic understanding to achieve substantial performance gains. However, these methods do not usually employ such BERT-like models to perform WSD directly. Instead, they strategically decompose the WSD problem into simpler subtasks or reformulate it into alternative paradigms (Huang et al., 2019; Yap et al., 2020; Barba et al., 2021b). This reduction in classification complexity enables the models to operate more effectively, further improving disambiguation accuracy. Moreover, a recent probing study (Teglia et al., 2025) suggests that BERT-like models can inherently distinguish word senses to some extent without explicit fine-tuning for WSD. This observation raises a critical question: *what factors subsequently affect a model's ability to further refine this inherent sense discrimination during supervised fine-tuning?*

We investigate this issue through three analytical lenses centered on the training corpus: **Distribution**, **Dosage**, and **Density**. **Distribution** refers to the frequency distribution of word senses within the corpus, particularly the long-tail phenomenon, in which a small number of senses are highly fre-

quent while many others are rare. A common hypothesis is that models trained on such imbalanced data tend to overfit frequent senses, resulting in degraded performance on rare ones (Blevins and Zettlemoyer, 2020; Su et al., 2022). To test this hypothesis, we artificially manipulate the sense distribution during training to examine whether the over-representation of high-frequency senses directly causes performance degradation on low-frequency ones. **Dosage** characterizes the effective scale of the training corpus in terms of unique, non-redundant training instances, analogous to the active ingredient in a medicinal dose, excluding repetitive data that the model has already learned. To quantify how dosage affects disambiguation learning, we propose the *Disambiguation Inflection Point (DIP)*, defined as the number of transitional layer from general semantic understanding to explicit sense disambiguation. By tracking how the DIP shifts across models trained with varying dosages, we analyze how this factor influences a model's learning of WSD. **Density** measures the density of disambiguating semantic cues, specifically, contextual semantic keywords that directly signal the correct sense of a polysemous word. Given the subjective nature of identifying such keywords, we focus on training instances whose senses are consistently predicted with either perfect accuracy (always correct) or zero accuracy (always incorrect). For this subset, we apply a two-stage annotation protocol, manually annotating the correct sense, perceived disambiguation difficulty and the most influential contextual keywords. These annotations enable us to examine whether poor performance on certain senses arises because key disambiguating cues are sparsely distributed within complex contexts, making it difficult for the model to attend to them effectively.

To summarize, we aim to address the following three research questions (RQs) in this paper:

- RQ1: How does distribution of senses in the training corpus influence a model's ability to learn word senses?
- RQ2: How does dosage of training corpus affect a model's acquisition of WSD capability?
- RQ3: How does density of contextual semantic keywords affect a model's learning to discriminate word senses?

Our findings reveal that while distributional imbalance does affect performance on certain senses, it is the absolute number of training instances per sense that emerges as a more fundamental determinant of success. Furthermore, we demonstrate that dosage significantly influences the model's internal processing: a larger dosage streamlines the semantic understanding process, allowing models to achieve better performance with fewer intermediate layers. Finally, we show that the density of semantic keywords critically impacts the model's attention allocation; contexts with sparse cues lead to attention being diverted to other uninformative tokens, thereby limiting disambiguation accuracy. By strategically increasing density of keywords, we can refocus the model's attention and unlock substantial performance gains on previously unlearned (accuracy is zero) senses.

2 Related Work

2.1 Fine-tuned BERT for WSD

With the rise of pre-trained language models and increased computational power, supervised learning has become the dominant paradigm in WSD (Bevilacqua et al., 2021). Most supervised approaches integrate additional semantic information, typically glosses, into the training process, which models can effectively exploit for disambiguation. For example, GlossBERT (Huang et al., 2019) reformulates WSD as a binary classification task by concatenating the context with the gloss of each candidate sense. Similarly, Yap et al. (2020) cast WSD as a gloss selection problem, where the model learns to choose the best contextgloss pair from a candidate set augmented with example sentences. Following the same line, ConSec (Barba et al., 2021b), ESC (Barba et al., 2021a), and KELESC (Zhang et al., 2022) formulate WSD as a text extraction problem: the input is enriched with glosses of neighboring words, and the model is required only to predict the start and end positions of the correct sense within the input. In addition, SensEmBERT (Scarlini et al., 2020a), ARES (Scarlini et al., 2020b), and SREF (Wang and Wang, 2020) also leverage gloss information to construct sense-aware representations for WSD. By explicitly providing concise definitions of candidate senses, these methods allow models to directly utilize gloss tokens for more effective disambiguation. At the same time, by reformulating the task, they avoid direct classification over a large

inventory of fine-grained sense labels, thereby substantially reducing task complexity.

SANDWiCH (Olivares et al., 2025) adopts a different strategy by employing a coarse sense retrieval module together with a semantic network, namely BabelNet (Navigli and Ponzetto, 2010), to progressively narrow down the candidate senses to a small set of the most plausible definitions. The backbone model then selects the correct sense from these pre-filtered context-definition pairs to complete the disambiguation task. To the best of our knowledge, SANDWiCH represents the current state of the art in WSD.

Collectively, these methods reduce both the burden on models to independently extract discriminative contextual features and the complexity of the classification space. As a result, they can more easily solve the assigned subproblems and, consequently, achieve more accurate WSD.

2.2 Approaches for Addressing Imbalanced Datasets

The power-law distribution is a pervasive characteristic of human language, with word frequency well described by Zipf's law (Zipf, 2016). Prior research has shown that this linguistic regularity also manifests in WSD corpora, particularly in SemCor (Miller et al., 1993). To mitigate the adverse effects of such imbalanced datasets, various methods have been proposed. For instance, BEM (Blevins and Zettlemoyer, 2020) introduces a dual-encoder framework that jointly embeds contexts and word senses into a shared vector space, while incorporating gloss information to enhance its learning for low-frequency senses. This approach significantly reduces error rates on rare senses. Similarly, the Z-reweighting strategy (Su et al., 2022) assigns frequency-aware weights to different senses, further improving model performance on infrequent senses. QiWSD (Zhang and Li, 2025) inspired by quantum theory, alleviates the dependence on big data and improving the accuracy of representations. Building on these efforts, our work investigates the specific impact of long-tail training distributions on a model's ability to learn and disambiguate word senses.

3 Method

In this section, we detail our methodology for investigating the influence of distribution, dosage, and density on a model's learning to WSD.

3.1 Distribution

A dataset exhibits a long-tail distribution when the vast majority of instances are concentrated in a small subset of label classes, while the remaining classes account for only a minor proportion of the data. Intuitively, we hypothesize that models trained on such imbalanced data tend to overfit high-frequency senses, achieving high accuracy on frequent senses but performing poorly on rare ones. This hypothesis is consistent with prior research (Postma et al., 2016), which identifies data distribution imbalance as one of the key factors influencing the performance of machine-learning-based WSD methods. SemCor (Miller et al., 1993), one of the most widely used datasets for supervised WSD, exemplifies this issue. Although it is a high-quality, human-annotated corpus, its source texts are drawn from real-world language, which naturally follows Zipf's law and thus exhibits a pronounced long-tail distribution of word senses (Su et al., 2022). Therefore, data *distribution* is the first factor we consider as potentially critical to a model's WSD capability.

To investigate this, we first train multiple language models M_i $(i = 1, \cdots, n)$ under identical training configurations and evaluate their per-sense disambiguation accuracy. For each sense s, let $A_{s,i} \in [0, 1]$ denotes the accuracy of model M_i on s. Based on the consistency of model performance across $\{M_i\}$, we group senses in the full inventory S and define four categories: (1) senses for which all models achieve perfect accuracy, denote as $S_{acc1} = \{s \in S \mid A_{s,i} = 1, \forall i\}$; (2) senses for which all models achieve zero accuracy, denote as $S_{acc0} = \{s \in S \mid A_{s,i} = 0, \forall i\}$; (3) senses on which all models achieve the same, non-extreme accuracy, denote as $S_{mid} = \{s \in S \mid A_{s,i} = \alpha_s, \alpha_s \in (0, 1), \forall i\}$; (4) senses whose performance varies across different training runs, denote as $S_{var} = \{s \in S \mid \exists i, j : A_{s,i} \neq A_{s,j}\}$. This categorization helps prevent the influence of data distribution on certain senses from being obscured by the overall imbalance of the dataset. To further disentangle the role of data imbalance, we construct a balanced dataset D_k by randomly sampling k instances per sense from the original dataset D (senses with fewer than k instances are removed). We then train a new model M^k on D_k and obtain its per-sense accuracy α_s^k.

Finally, we analyze the performance shift for senses in each category and use Mean Signed Dif-

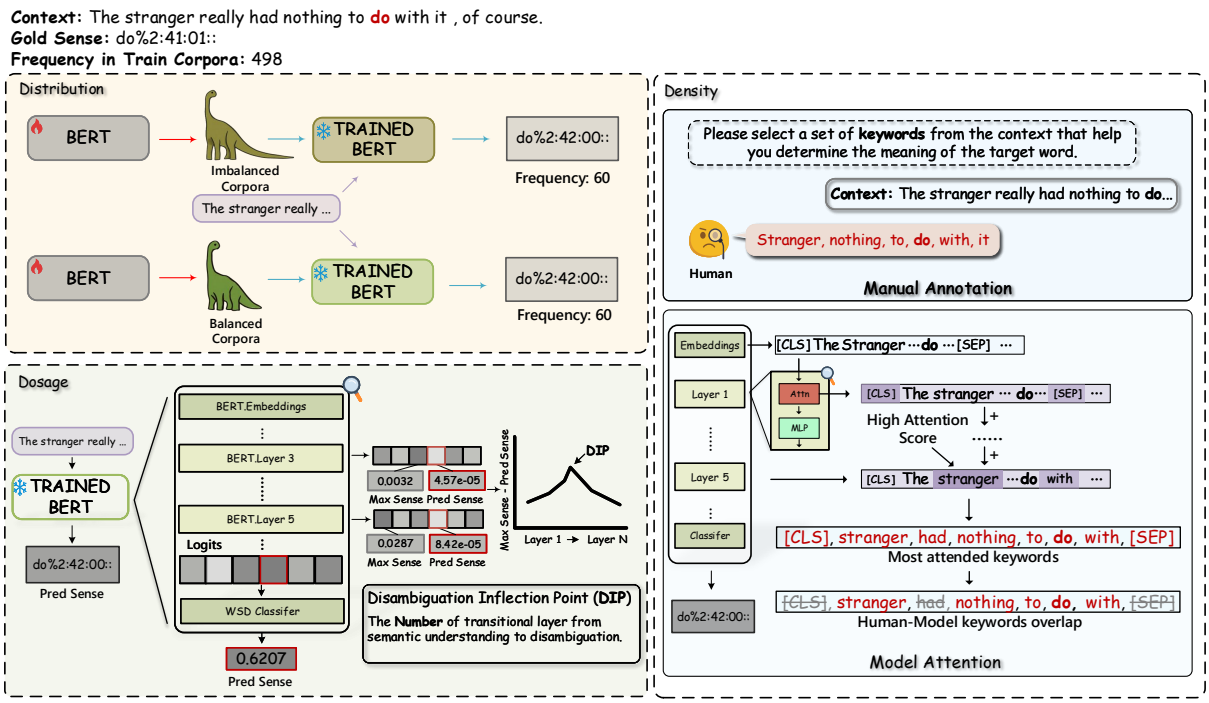

Figure 1: Methodology Overview: (1) **Distribution**: compare disambiguation performance across sense-frequency categories under different label distributions; (2) **Dosage**: track DIP variation across increasing data dosages; (3) **Density**: identify the tokens most attended by the model and evaluate their alignment with human-annotated keywords.

ference (MSD) to quantify both the magnitude and direction of change:

$$\mathrm{MSD}(S') = \frac{1}{|S'|} \sum_{s \in S'} \left(\alpha_s^k - \frac{1}{N} \sum_{i=1}^{N} \alpha_{s,i} \right), \quad (1)$$

where $S' \in \{S_{acc1}, S_{acc0}, S_{mid}, S_{var}\}$.

3.2 Dosage

The term *dosage* is commonly used in medicine to denote the amount of an active ingredient contained in a formulation. By analogy, we define dosage as the "active ingredient" of the data that a model is actually exposed to, specifically, the amount of unique, non-repetitive training instances encountered throughout the supervised fine-tuning phase. The magnitude of dosage reflects the volume of WSD-specific knowledge available to the model, and thus constitutes an important factor that influences its ability to learn and discriminate word senses.

3.2.1 Disambiguation Inflection Point (DIP)

Consider a BERT-like model with n intermediate layers, trained to disambiguate among m candidate word senses. Typically, the output of the final layer before the classification head is referred to as the logits vector $\mathbf{L} = [l_1, l_2, \ldots, l_m]$, which represents the model's raw prediction scores. The final predicted sense is obtained as $s^{pred} = \mathrm{Classifier}_S(\mathbf{L})$. If we instead feed the output of the ℓ-th intermediate layer l_ℓ ($1 \le \ell \le n$) into the same classifier $\mathrm{Classifier}_S$, we obtain a layer-wise sense prediction s_ℓ^{pred} and corresponding index $\widehat{s}$. The corresponding logits vector for l_ℓ is denoted $\mathbf{L}_\ell = [l_{1,\ell}, l_{2,\ell}, \ldots, l_{m,\ell}]$, with $\mathbf{L}_n$ being the final-layer logits.

Let $l_{pred,\ell} = l_{\widehat{s},\ell}$ denote the logit score of the final predicted sense at layer l_ℓ. Then we define the Prediction Deviation at layer l_ℓ as:

$$\mathrm{PD}_\ell = \left| l_{\mathrm{pred},\ell} - l_{\mathrm{pred},n} \right|. \quad (2)$$

The sequence $\{\mathrm{PD}_\ell\}_{\ell=1}^{n}$ satisfies two properties:

$$\mathbf{P}_1 : \mathrm{PD}_n = 0 \; ; \; \{\mathrm{PD}_\ell\}_{\ell=1}^{n-1} \ge 0.$$
$$\mathbf{P}_2 : \exists\, p \in [1, n-1], \forall\, p < j \le n : \mathrm{PD}_j \le \mathrm{PD}_p.$$

We finally define p as the *Disambiguation Inflection Point* (DIP). And, layers L_ℓ with $\ell \le p$ are interpreted as being primarily responsible for contextual semantic understanding, whereas layer L_ℓ with $\ell > p$ are considered to engage in explicit sense discrimination.

To study the influence of dosage on WSD learning, we train multiple BERT-like models with training sets of varying dosage and analyze how the DIP shifts across these models.

3.3 Density

When humans disambiguate the sense of a polysemous word in a sentence, we typically rely on certain contextual semantic keywords that indicate the intended meaning. For example, in "*I eat an apple*," the verb *eat* signals that *apple* refers to the fruit rather than the technology brand. However, in more complex contexts (see examples in Appendix D), such semantic cues may be subtle or ambiguous, making the correct sense more difficult to identify. We define *density* as the density of disambiguating cues in the surrounding context. In practice, we approximate density by computing the proportion of manually identified keywords within the context. To study its effect on WSD learning, we employ a two-stage annotation protocol to label these keywords.

3.3.1 Annotation of Semantic Keywords

To ensure that the annotated contextual cues are meaningful for model behavior analysis, we focus on training instances from the sense categories S_{acc0} and S_{acc1}. From these categories, we randomly sample 100 instances for manual annotation. Each annotation record contains three components: (1) the Correct Sense, (2) the Perceived Disambiguation Difficulty, and (3) the Contextual Semantic Keywords. For the keywords, annotators are instructed to select, in descending order of importance, up to three keywords in the context that most strongly influence the sense of the target word. The upper limit of three is imposed because, as importance decreases, the boundary between relevant and marginal cues becomes increasingly ambiguous, which may otherwise introduce noise into the annotations.

In the second stage, to investigate the effect of density on WSD learning, we randomly sample 150 instances from S_{acc0} and focus exclusively on annotating the contextual semantic keywords. Unlike the first stage, annotators are not required to provide keywords in ranked order, and no upper limit is imposed on the number of keywords. The resulting keyword sets are later used to replace the corresponding context tokens in the training data, allowing us to examine how increasing density influences the model's WSD capability. All annotations are conducted by ourselves rather than external linguistic experts, in order to avoid potential underestimation of disambiguation difficulty and overly conservative keyword selection by domain specialists. Further annotation details and the annotation interface are provided in Appendix B.

3.3.2 Identification of Semantic Keywords from Model Attention

To identify the contextual cues that the model relies on during sense disambiguation, we analyze the attention patterns of its intermediate layers. Consider a BERT-like model with n intermediate layers and m attention heads per layer, and an input token sequence $\mathbf{t} = [t_1, t_2, \cdots, t_{target}, \cdots, t_T]$. Let the attention matrix of the j-th head in the i-th layer be denoted by $\mathbf{A}^{(i,j)} \in \mathbb{R}^{T\times T}$, where T is the sequence length. The attention paid by the target token at position pos_{target} to the token at position p in layer i is computed as:

$$\text{att}_p^{(i)} = \sum_{j=1}^{m} \mathbf{A}^{(\mathbf{i},\mathbf{j})}[\text{pos}_{target}, p]. \quad (3)$$

The *Total Attention* (TotalAtt) assigned to token p over the entire disambiguation process is then defined as the sum over all intermediate layers[1]:

$$\text{TotalAtt}_p = \sum_{i=1}^{n} \text{att}_p^{(i)}. \quad (4)$$

We then rank all contextual tokens according to their TotalAtt scores, and denote the top-k most attended disambiguation keywords as K_k.

To investigate whether performance failures stem from missing key semantic cues, as well as the degree to which the model's attended keywords agree with human-annotated ones, we compare K_k with the human-annotated keywords. We quantify their alignment using a *Weighted Hit Rate* (WHR) for the i-th human-annotated keyword:

$$\text{WHR}_i@k = \frac{1}{N}\sum_{n=1}^{N} \frac{\max\big(0,\ k - r_i^{(n)} + 1\big)}{k}, \quad (5)$$

where N is the total number of instances, and $r_i^{(n)}$ denotes the rank at which the model attends to the

[1] If a word is tokenized into multiple subwords by the tokenizer, we average the attention scores across its subword tokens to obtain a single score for the original word.

i-th annotated keyword in the n-th instance. A higher value indicates stronger agreement between model attention and human focus.

In addition, we leverage Normalized Discounted Cumulative Gain (nDCG) as a metric to evaluate the model's overall attentional focus on all annotated keywords:

$$\text{nDCG@}k = \frac{\sum_{i=1}^{k} \frac{2^{rel_i}-1}{\log_2(i+1)}}{\sum_{i=1}^{k} \frac{2^{rel_i^{ideal}}-1}{\log_2(i+1)}} \quad (6)$$

where rel is the relevance score and rel^{ideal} denotes the ideal results sorted in descending order of relevance scores.

4 Experimental Settings

4.1 Datasets

We use two public WSD training corpora (SemCor and OMSTI) and five public evaluation benchmarks (ALL).

- **SemCor** (Miller et al., 1993): The most commonly used human-annotated corpus. We additionally construct a balanced subset **SemCor-5** by sampling five instances per sense; and a sub dataset **SemCor-sub** with the original frequency distribution. **SemCor-key** is obtained by replacing the original context of 150 instances in SemCor with their human-annotated keywords.
- **OMSTI** (Taghipour and Ng, 2015): An automatically annotated corpus with 3,730 senses and 911,134 instances.
- **ALL** (Raganato et al., 2017): A consolidated evaluation set combing five benchmarks: SemEval-2007, SemEval-2013, SemEval-2015, Senseval-2, and Senseval-3.

4.2 Models

We primarily probe **BERT-base-uncased** (Devlin et al., 2019), a transformer encoder pre-trained on large-scale English text and widely used in supervised WSD.

4.3 Experimental Details

All experiments are conducted on two NVIDIA GeForce RTX 3090 GPUs. We largely follow the original training configurations without extensive hyperparameter tuning, as our goal is to analyze model behavior rather than optimize performance.

For **BERT-base-uncased**, we use: batch size 256, learning rate 1×10^{-4}, 10 epochs, weight decay 0.01, 12 layers, hidden size 768, intermediate size 3,072, and 12 attention heads.

4.4 Metrics

Because several experiments involve restricted or balanced subsets and our focus is on factors' effects, we report *restricted* micro-F1 (m-F1) and macro-F1 (M-F1), excluding senses that never appear in the training data from evaluation.

5 Results and Analysis

We present experimental results to illustrate the influence of the three factors, *Distribution*, *Dosage*, and *Density*, on a model's learning to disambiguate word senses.

5.1 Impact of Distribution

Sense Sets	MSD	ΔN	ΔM-F1
S_{acc1}	-0.0636	-22.94	-0.2031
S_{acc0}	0.0294	-1.55	**0.2829**
S_{mid}	0.0002	-1.78	-0.0120
S_{var}	-0.0219	-10.55	-0.0093

Table 1: MSD (Mean Signed Difference), ΔN (Mean change in number of data instances) and ΔM-F1 (Difference in macro-F1) of three BERT models trained on SemCor and one on SemCor-5 across four sense sets.

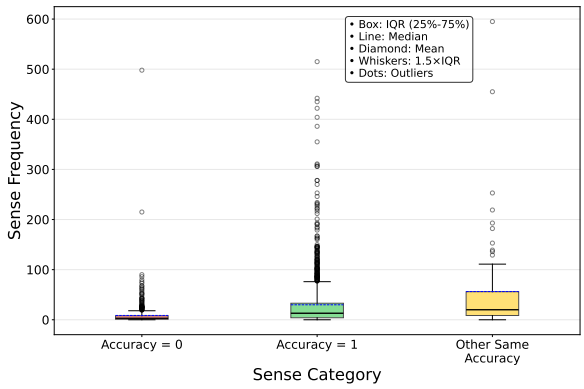

Figure 2: Boxplot of the per-sense training instance counts across three sense categories (S_{acc0}, S_{acc1}, S_{mid}). For readability, only outliers with counts below 1,000 are retained.

Table 1 reports the Mean Signed Difference (MSD), mean change in number of instances per sense (ΔN), and ΔM-F1 for the four sense

categories S_{acc0}, S_{acc1}, S_{mid}, and S_{var} (cf. Section 3.1), using BERT models[2] trained on SemCor and SemCor-5. The positive MSD and ΔM-F1 on S_{acc0} indicate that the imbalanced distribution does hinder performance on certain senses. However, the relatively small change in accuracy across all categories ($|\text{MSD}| < 0.1$) suggests that modifying the data distribution alone has only a limited effect on the model's ability to learn word senses. To demonstrate the validity of the experimental results, we conduct a significance test on the MSD. The results show that all four sense categories exhibit statistically significant MSD values (p-value < 0.05), indicating that the shift is consistent across runs and unlikely to be due to random variation (Appendix A.2). Moreover, the MSD and ΔN values for S_{acc1}, S_{mid}, and S_{var} show that performance generally degrades when the number of training instances decreases, regardless of distributional balance. This pattern suggests that the absolute amount of data available per sense, rather than its relative frequency within the dataset, may be the dominant factor driving performance differences. To further examine this, we plot the per-sense instance counts in SemCor for S_{acc1}, S_{mid}, and S_{var}. As shown in Figure 2, senses in S_{acc0} indeed contain substantially fewer instances than those in the other categories.

Therefore, for **RQ1**, we conclude that while dataset imbalance can negatively affect the disambiguation accuracy of certain senses, the insufficient absolute number of training instances per sense is likely a key contributing factor to poor performance on these senses.

5.2 Impact of Dosage

Datasets	Dosage	DIP	m-F1	M-F1	Pred Prob
SemCor-sub	44,554	11	0.60	0.56	0.69
SemCor	226,036	6	0.70	0.60	0.85
OMSTI-30	300,000	5	0.66	0.52	0.93
OMSTI-40	400,000	4	0.69	0.55	0.95
OMSTI-80	800,000	4	0.67	0.51	0.91

Table 2: DIP, Micro-F1 (m-F1), Macro-F1 (M-F1) and prediction probability (Pred Prob) for BERT models trained with different data dosages.

Table 2 reports the Disambiguation Inflection Point (DIP) of BERT models trained under different data dosages. We observe a clear trend that the

[2]Due to the reduced training size, the number of training epochs on SemCor-5 is increased to 20.

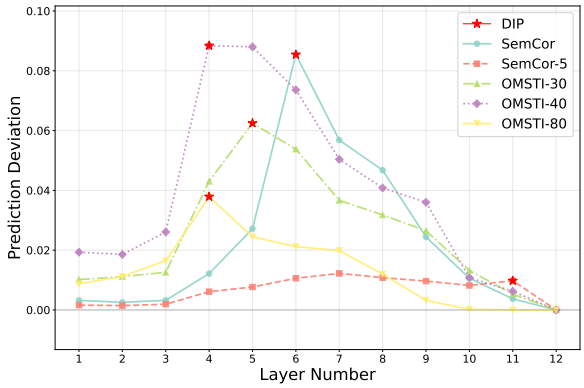

Figure 3: Prediction Deviation per layer for BERTbase models trained with different dosages. The red pentagram marks the Disambiguation Inflection Point (DIP).

DIP decreases as dosage increases. This suggests that exposure to more unique training instances enables the model to acquire richer WSD-relevant knowledge, thereby requiring fewer intermediate layers for contextual semantic understanding. As a result, the remaining layers can be devoted more explicitly to sense disambiguation. Consistently, a larger dosage also improves the model's prediction confidence, as reflected by the increase in averaged prediction probability (Pred Prob). Figure 3 further illustrates the per-layer Prediction Deviation (PD) for models trained with different dosages. For the model trained on OMSTI-80, both PD_{11} and PD_{12} (i.e., the deviations at the 11-th and 12-th layers; cf. Section 3.2.1) are zero, indicating that disambiguation has already been completed by the 11-th layer. This finding confirms that only a limited number of intermediate layers are necessary for sense separation. Motivated by this observation, we train an 11-layer BERT model on OMSTI-80. As shown in Table 3, the model not only improves disambiguation performance but also reduces training time by 7.79%.

Taken together, for **RQ2** we conclude that data

Datasets	Layers	m-F1	M-F1	Training Time
OMSTI-80	12	0.67	0.51	5:15h
	11	**0.69**	**0.55**	**4:50h**

Table 3: Micro-F1 (m-F1), Macro-F1 (M-F1), and Training Time (hours:minutes) for BERT models with different numbers of layers trained on OMSTI-80.

dosage affects both (i) the number of layers required for contextual understanding and (ii) the confidence of the final sense prediction. Higher dosage leads to higher disambiguation confidence

and fewer layers allocated to semantic comprehension. Moreover, the fact that a finite subset of intermediate layers suffices for disambiguation suggests that, in principle, BERT-like models with fewer total layers can be employed when the training data dosage is sufficiently large. Additional experiments with RoBERTa and models of varying depths are provided in Appendix A.3.

5.3 Impact of Density

Train Corpus	Sense Set	Weighted Hit Rate			
		WHR_1	WHR_2	WHR_3	nDCG
SemCor	S_{acc0}	0.302	0.230	0.157	0.304
	S_{acc1}	0.354	0.293	0.154	0.305
SemCor-key	S_{acc0}	0.366	0.270	0.230	0.375
	S_{acc1}	0.421	0.310	0.240	0.380

Table 4: Weighted Hit Rate (WHR) and Normalized Discounted Cumulative Gain (nDCG) of the top-10 tokens attended by the BERT model on human-annotated keywords. The evaluation is conducted on the annotated portion of SemCor.

Table 4 reports the alignment between the top-10 semantic keywords attended by the BERT model during WSD and those annotated by humans. We observe that keywords judged by humans as more important for disambiguation achieve higher WHR and nDCG, indicating a certain degree of consistency between human focus and model attention. However, we also find that the tokens receiving the highest attention are predominantly special tokens that frequently appear in the context (see Appendix C), causing the model's attention to be generally under-allocated to meaningful disambiguation keywords (Clark et al., 2019). Furthermore, when the model is trained on the variant where the original context of 150 instances is replaced with their human-annotated keywords (see examples in Appendix D), its attention to contextual keywords increases substantially. Consistently, Table 5 shows that the model trained on SemCor-key achieves significantly better performance on the corresponding senses than the model trained on the original SemCor, with m-F1 increasing from 0 to 0.34 and M-F1 from 0 to 0.16. These results indicate that increasing the keyword density in the training data strengthens the model's attention on key contextual cues and improves its WSD ability.

Test Set	Train Corpus	m-F1	M-F1
ALL(full set)	SemCor	0.7091	0.5882
	SemCor-key	**0.7175**	**0.6009**
ALL(replaced only)	SemCor	0.0	0.0
	SemCor-key	**0.3396**	**0.1550**

Table 5: Models trained on the original SemCor and on its keyword-replaced variant (SemCor-key) are evaluated using Micro-F1 (m-F1) and Macro-F1 (M-F1). ALL (replaced only) reports results only on senses whose instances are modified in SemCor-key.

In summary, for **RQ3**, complex contexts can reduce the attention allocated to contextual semantic keywords, thereby limiting disambiguation performance. Increasing keyword density in the training data mitigates this effect by directing more attention toward informative semantic cues, ultimately enhancing the model's WSD capability.

6 Conclusion

We present a probing study that investigates the factors dominating a BERT-like model's acquisition of Word Sense Disambiguate (WSD) during supervised fine-tuning. In particular, we examine three factors: **Distribution**, **Dosage**, and **Density**, and analyze their respective influences on sense learning behavior. Our results show that **distribution** does affect model performance on certain senses; however, the absolute number of training instances per sense emerges as another key factor underlying performance variation. **Dosage**, in contrast, has a substantial impact on the model's internal processing dynamics: a larger dosage reduces the number of intermediate layers required for semantic understanding and increases the model's disambiguation confidence. On high-dosage datasets, models with fewer intermediate layers can be trained effectively, achieving better disambiguation performance while also reducing training time. We further find that when the **density** of semantic cues, specifically contextual semantic keywords, in the training context is low, the model tends to assign relatively low attention ranks to these keywords, which limits its disambiguation ability. By replacing the original context of 150 training instances with their manually annotated keywords, the micro-F1 score on the corresponding senses increases from 0 to 0.34, and the macro-F1 score increases from 0 to 0.16. These findings indicate that increasing the density of disambiguating cues in the training data encourages the model to better attend to key contextual information, thereby improving its WSD performance.

Limitations

Our experiments do not cover all BERT-like architectures, and the influence of the three investigated factors may vary substantially across different model families. We also observe that even models with identical architectures (e.g., BERT and RoBERTa) may exhibit markedly different behaviors due to differences in pre-training strategies. Therefore, our findings should be understood as reflecting the most general effects of these factors, and further work is needed to conduct more fine-grained analyses across a broader range of models.

In addition, we approximate the density of disambiguating cues using the proportion of annotated keywords in the context, which is admittedly a coarse approximation. To validate its effect, we adopt an extreme intervention strategy by fully replacing the original context of selected instances with their annotated keyword sets. Although this intervention leads to improved performance on the corresponding senses, it inevitably removes other semantic information from the context. Moreover, the identification of contextual keywords currently relies entirely on human judgment, meaning that not all replacements are guaranteed to be beneficial to the model, and only 150 instances can be annotated within our current capacity. Future work is therefore required to develop more reliable keyword extraction and replacement strategies, enabling larger-scale and less subjective dataset refinement.

References

Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021a. ESC: Redesigning wsd with extractive sense comprehension. In *Proceedings of the 2021 Conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672.

Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2021b. ConSeC: Word sense disambiguation as continuous sense comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1503.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, pages 4330–4338.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017.

Yang Chen, Vedaant Shah, and Alan Ritter. 2025. Translation and fusion improves cross-lingual information extraction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 7744–7764.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of berts attention. In *Proceedings of the 2019 ACL workshop BlackboxNLP: analyzing and interpreting neural networks for NLP*, pages 276–286.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuan-Jing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3509–3514.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

George A. Miller. 1994. WordNet: A lexical database for English. In *Proceedings of the 1994 Human Language Technology Workshop Held at Plainsboro*.

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *Proceedings of the 1993 Human Language Technology Workshop Held at Plainsboro*.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225.

Daniel Guzman Olivares, Lara Quijano, and Federico Liberatore. 2025. Sandwich: Semantical analysis of neighbours for disambiguating words in context ad hoc. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 7019–7033.

Marten Postma, Rubén Izquierdo, and Piek Vossen. 2016. More is not always better: Balancing sense distributions for all-words word sense disambiguation. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3496–3506.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 99–110.

Abudurexiti Reheman, Hongyu Liu, Junhao Ruan, Abudukeyumu Abudula, Yingfeng Luo, Tong Xiao, and Jingbo Zhu. 2025. Enhancing neural machine translation through target language data: A kNN-LM approach for domain adaptation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 10053–10065.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020a. Sensembert: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8758–8765.

Bianca Scarlini, Tommaso Pasini, Roberto Navigli, and 1 others. 2020b. With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539.

Ying Su, Hongming Zhang, Yangqiu Song, and Tong Zhang. 2022. Rare and zero-shot word sense disambiguation using z-reweighting. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 4713–4723.

Kaveh Taghipour and Hwee Tou Ng. 2015. One million sense-tagged instances for word sense disambiguation and induction. In *Proceedings of the 19th Conference on Computational Natural Language Learning*, pages 338–344.

Simone Teglia, Simone Tedeschi, and Roberto Navigli. 2025. How much do encoder models know about word senses? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 2266–2277.

Ming Wang and Yinglin Wang. 2020. A synset relation-enhanced framework with a try-again mechanism for word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6229–6240.

Lingxiao Wei, He Yan, Lu Xiangju, Junmin Zhu, Jun Wang, and Wei Zhang. 2025. CNNSum: Exploring long-context summarization with large language models in chinese novels. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8034–8062.

Boon Peng Yap, Andrew Koh, and Eng Siong Chng. 2020. Adapting BERT for word sense disambiguation with gloss selection objective and example sentences. In *Findings of the Association for Computational Linguistics: ACL 2020*, pages 41–46.

Guobiao Zhang, Wenpeng Lu, Xueping Peng, Shoujin Wang, Baoshuo Kan, and Rui Yu. 2022. Word sense disambiguation with knowledge-enhanced and local self-attention-based extractive sense comprehension. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4061–4070.

Junwei Zhang and Xiaolin Li. 2025. Quantum-inspired non-homologous representation constraint mechanism for long-tail senses of word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 25877–25885.

George Kingsley Zipf. 2016. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio books.

A Additional Experiments

This section reports additional analyses, including a discussion of the long-tail issue in the training data, experiments on RoBERTa (Liu et al., 2019), and further exploration of the effect of intermediate layer depth.

A.1 Sense Accuracy vs. Sense Frequency

Figure 4 shows the sense-level accuracy of BERT trained on SemCor, along with a curve indicating the corresponding sense frequency in the training set. We observe no obvious correlation between sense frequency and disambiguation accuracy. In fact, even without considering distributional imbalance, SemCor contains on average only 6.78 training instances per sense. This suggests that, rather than long-tail imbalance itself, the more fundamental issue is the overall insufficiency of data per sense. This hypothesis is further supported by the experiments in the next subsection.

A.2 Significance Tests of MSD

To verify the significance of the MSD, we conduct five independent training runs with different random seeds for both the original and balanced datasets, and performed statistical significance testing on the MSD values for each sense category. As shown in the table 6, all four sense categories exhibit statistically significant MSD values ($p < 0.05$), with particularly strong effects for S_{acc0}, S_{acc1}, and S_{var} ($|d| > 4$). Even for S_{mid}, where the absolute MSD is as small as 0.0002, the effect is statistically significant with a large effect size ($d = 1.42$), indicating that the shift is consistent across runs and unlikely to be due to random variation. These results confirm that the observed differences in disambiguation performance between the original and balanced training conditions are robust and statistically reliable.

A.3 Impact of Intermediate Layer Depth on DIP

In this subsection, we study how the Disambiguation Inflection Point (DIP) changes when the number of intermediate layers varies under the same data dosage. We additionally conduct experiments using RoBERTa.

RoBERTa-base-uncased (Liu et al., 2019) is an optimized variant of BERT with improved training strategies and larger-scale pre-training data. Specifically, we use a batch size of 128, a learning rate of 6×10^{-5}, 10 training epochs, weight decay of 0.1, 12 layers, hidden size 768, intermediate size 3,072, and 12 attention heads. The AdamW optimizer is configured with a warm-up ratio of 0.06, $\epsilon = 1 \times 10^{-8}$, $\beta_1 = 0.9$, and $\beta_2 = 0.98$.

Table 7 reports the DIP values of BERT and RoBERTa models with different layer depths. For BERT, varying the number of layers under the same dosage generally does not change the DIP, and the 11-layer model exhibits a lower DIP than the other configurations. This indicates that, for this 11-layer model, fewer layers are sufficient for semantic understanding, suggesting better internalization of semantic knowledge and improved WSD performance.

In contrast, RoBERTa consistently shows a DIP at the second-to-last layer, accompanied by abnormally low M-F1 and prediction probability. This behavior resembles that of BERT trained on the small SemCor-5 subset, leading us to hypothesize that the abnormal behavior of RoBERTa is caused by insufficient training data. To verify this, we train a 12-layer RoBERTa model on OMSTI-80 and observe that the DIP shifts to the third-to-last layer, while both M-F1 and prediction probability improve substantially.

These findings suggest that RoBERTa requires a larger training dosage than BERT to achieve stable and effective disambiguation behavior. Although tuning optimization parameters may partially compensate for this, it may still be difficult to reach optimal training outcomes under limited data. This further supports our earlier observation that, despite containing over 220K instances, SemCor still suffers from insufficient data per sense due to its large number of sense categories.

B Annotation Detail

This section describes the two-stage annotation interface and the associated annotation procedures.

B.1 The First Stage

Figures 5-7 illustrate the first-stage interface, where annotators (i) select the correct sense, (ii) rate the perceived disambiguation difficulty, and (iii) identify up to three key disambiguating keywords. Table 8 reports the statistics of the annotated data. Compared with instances in S_{acc1}, those in S_{acc0} exhibit lower HumanGold Agreement and higher perceived difficulty, suggesting that the model's difficulty perception is generally

Sense Set	MSD	SD	95% CI	t-value	df	p-value	Cohen's d
S_{acc0}	0.0294	0.0027	[0.0260, 0.0328]	24.10	4	<0.001	10.76
S_{acc1}	-0.0636	0.0030	[-0.0675, -0.0598]	-47.58	4	<0.001	-21.28
S_{mid}	0.0002	0.0001	[0.00003, 0.00037]	3.18	4	0.033	1.42
S_{var}	-0.0219	0.0052	[-0.0286, -0.0152]	-9.37	4	<0.001	-4.19

Table 6: The mean MSD, standard deviation (SD), 95% confidence interval (CI), paired t-test results, degree of freedom(df), and Cohen's d effect size for each sense set.

Dataset	**Model**	**Layers**	**DIP**	**M-F1**	**Pred Prob**
SemCor	BERT-base	9	6	0.59	0.826
		10	6	0.60	0.834
		11	5	0.61	0.832
		12	6	0.60	0.851
	RoBERTa-base	9	8	0.37	0.676
		10	9	0.37	0.575
		11	10	0.37	0.648
		12	11	0.36	0.687
SemCor-5	BERT-base	12	11	0.62	0.762
OMSTI-80	RoBERTa-base	12	10	0.53	0.954

Table 7: DIP, Macro-F1 (M-F1), and averaged prediction probability (Pred Prob) of BERT-base and RoBERTa-base models trained with different layer depths and datasets.

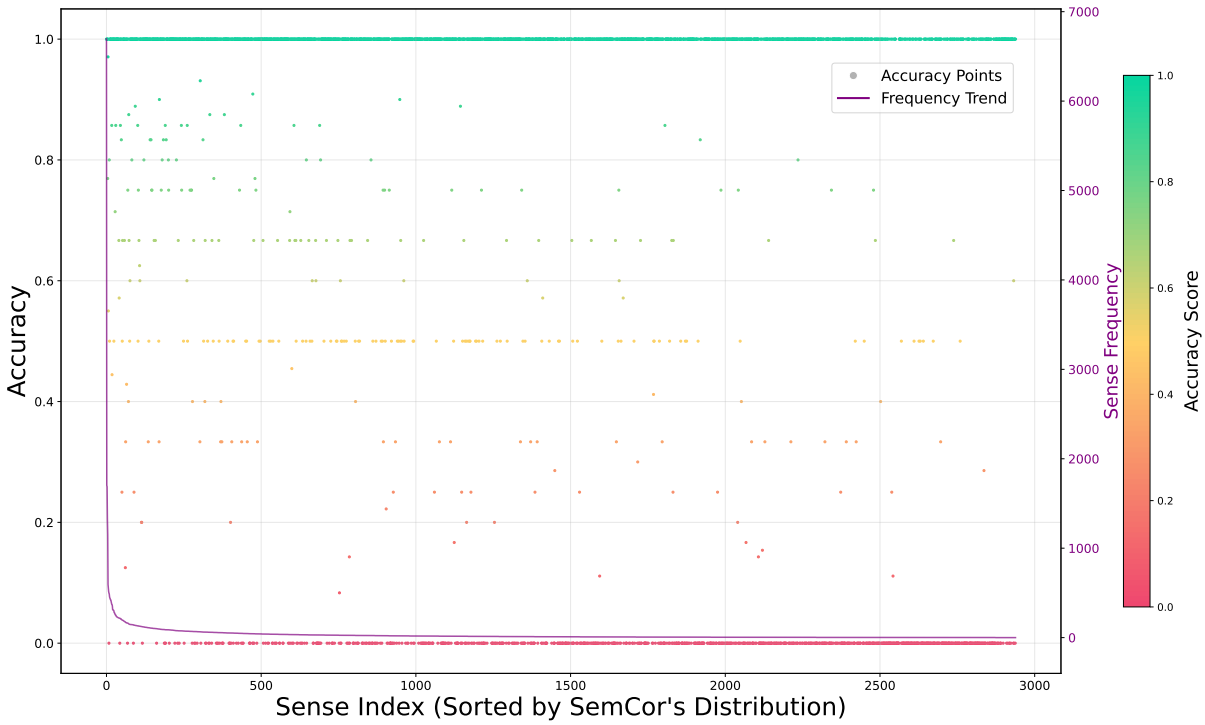

Figure 4: Accuracy of BERT trained on SemCor across word senses, sorted by frequency in SemCor.

consistent with that of human annotators.

We invite five colleagues to label the same set of instances. For each instance, we retain the annotation outcome with the highest agreement among annotators, and discard cases with highly disagreement (e.g., instances where more than two sense candidates are selected among the four options).

B.2 The Second Stage

Figure 8 shows the interface used in the second-stage annotation. In this stage, annotators focus exclusively on identifying the contextual keywords that most strongly influence the sense of the target word, without any constraint on the number of keywords or their ordering. Since keyword identification is highly subjective by nature, the 150 instances are evenly distributed among all annotators for one-pass annotation, and we do not enforce agreement across annotators in this stage.

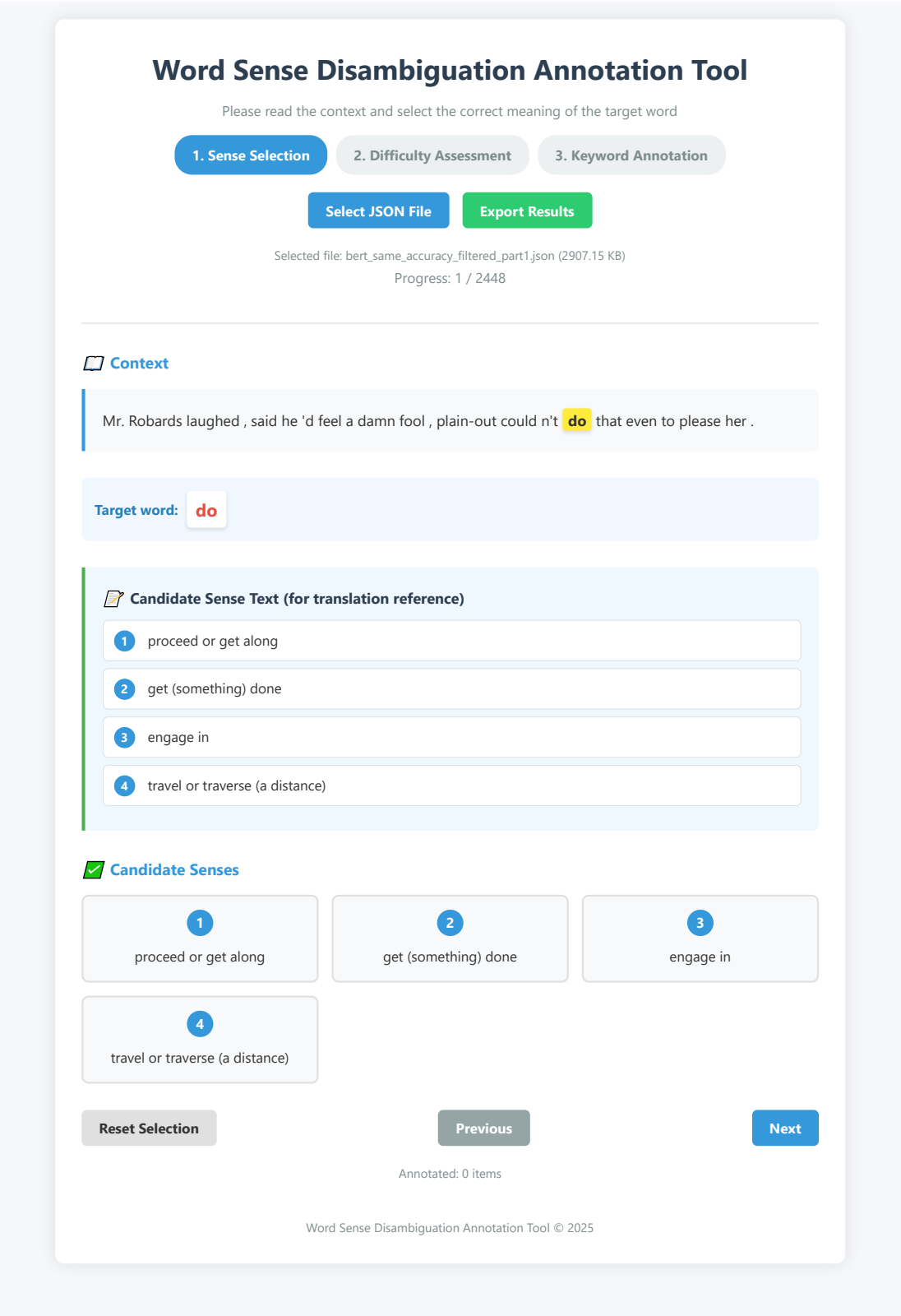

Figure 5: Interface for annotating correct word sense in the first stage of annotation.

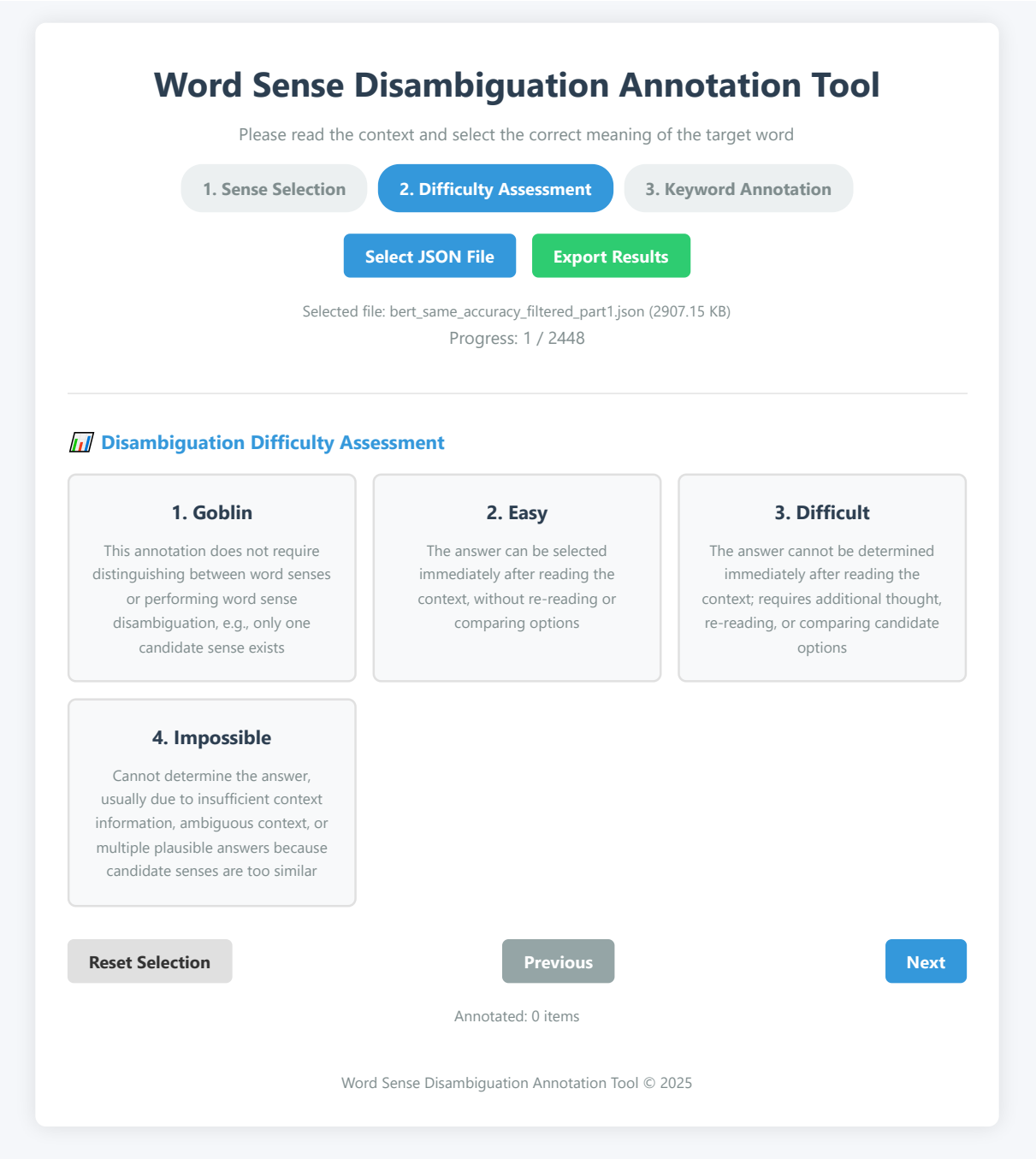

Figure 6: Interface for annotating disambiguation difficulty in the first stage of annotation.

Sense Category	Count	Human-Gold Agreement (%)	Avg. Perceived Difficulty
S_{acc0}	21	55.17%	2.690
S_{acc1}	79	74.65%	2.282

Table 8: Statistics of the 100 annotated instances in Stage-1 annotation. Human-Gold Agreement (%) denotes the proportion of user selections consistent with the gold label, and Avg. Perceived Difficulty indicates the average difficulty score assigned by annotators.

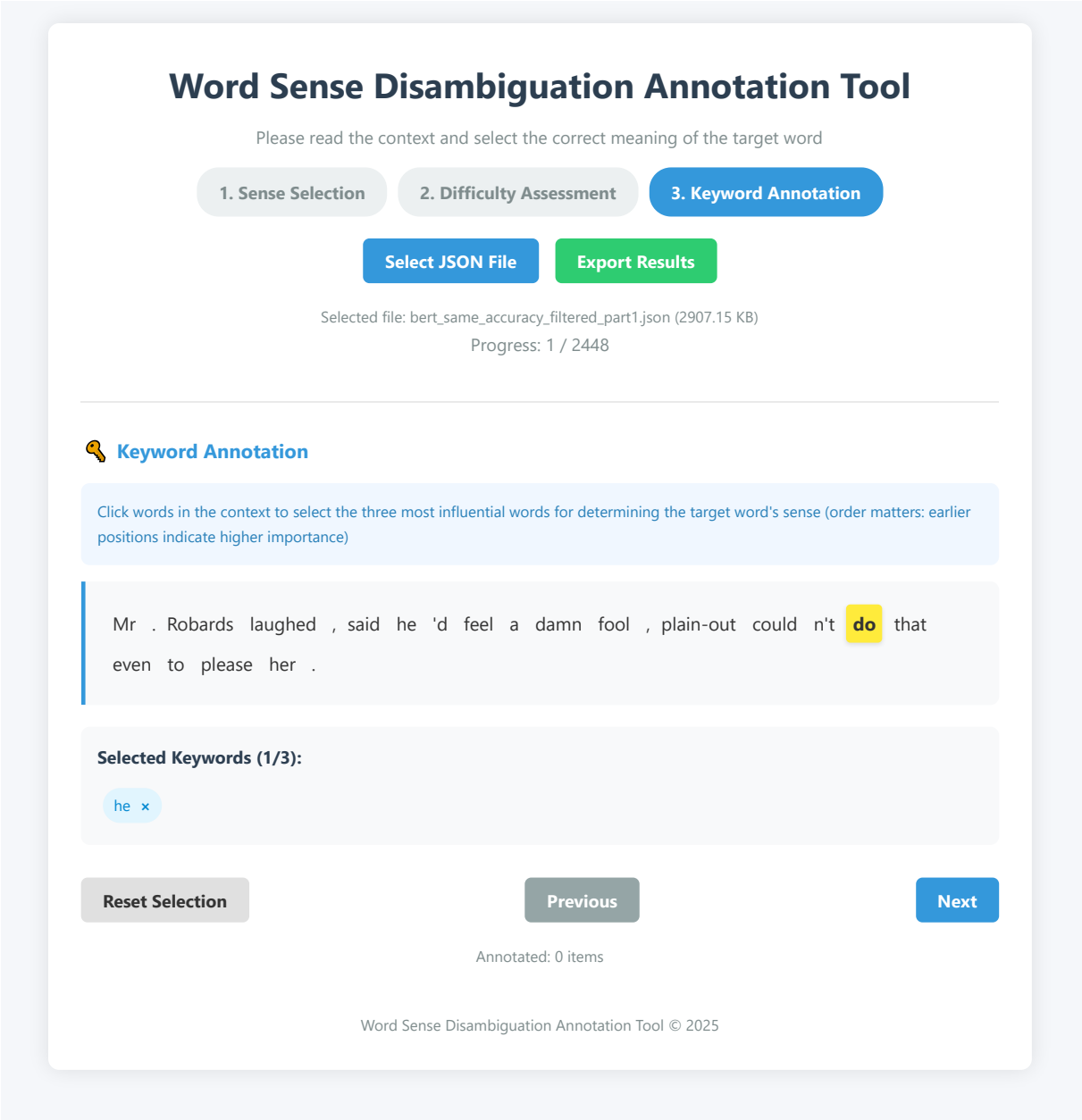

Figure 7: Interface for annotating keywords in the first stage of annotation.

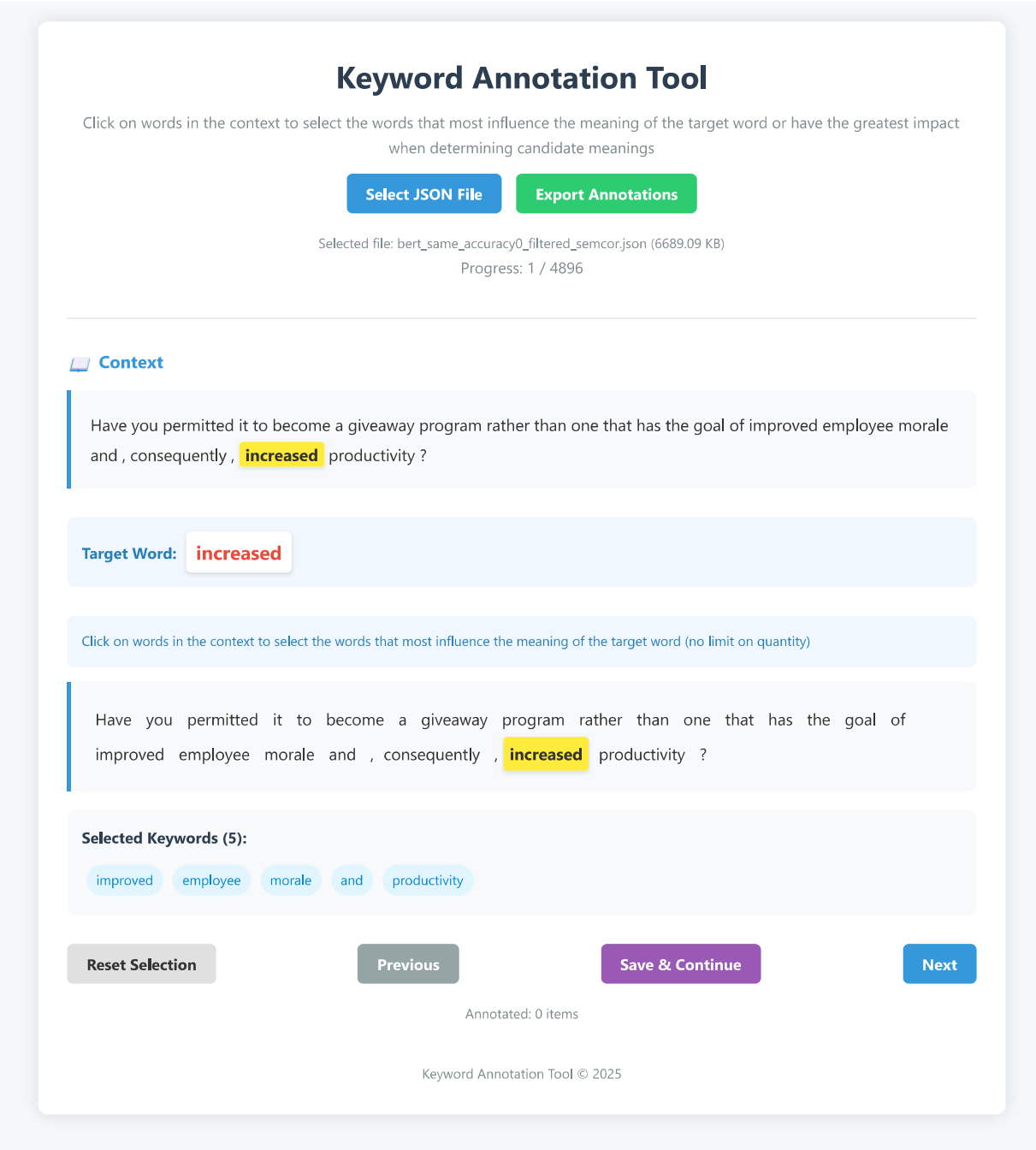

Figure 8: Interface for annotating keywords in the second stage of annotation.

C Overfitting of Attention to Special Tokens

In this section, we analyze an observation that the model assigns disproportionately high attention scores to special tokens. As shown in Table 9, the BERT model fine-tuned on SemCor tends to strongly attend to tokens that always appear in the input, such as `[CLS]`, `[SEP]`, `[TGT]`, and `[/TGT]`, where `[TGT]` and `[/TGT]` are task-specific markers indicating the target word span. We further find that the period symbol, which frequently occurs in the context, also ranks highly in attention.

After training on the keyword-replaced data, however, the attention rank of the period token decreases substantially, as punctuation is generally not included in the annotated keyword sets. This pattern suggests that the fine-tuned model exhibits a form of **attention overfitting**: during prediction, a considerable portion of attention is allocated to special or frequent tokens, which reduces the effective attention available for informative contextual cues and ultimately limits WSD performance.

	Special Tokens						
Train Corpus	[CLS]	[SEP]	[PAD]	[TGT]	[/TGT]	,	.
SemCor	2.5	1.2	17.8	7.1	5.1	13.5	3.5
SemCor-key	2.3	1.2	17.8	6.0	4.2	14.5	12.6

Table 9: Averaged attention ranking of special tokens when BERT performs WSD. "," denotes the comma token, and "." denotes the period token.

D Examples

This section presents representative examples from the training data, illustrating (i) simple contexts (all from S_{acc1}), (ii) complex contexts (S_{acc0}), and (iii) contexts rewritten using the manually annotated keyword sets (S_{acc0}).

Simple Contexts.

1. *What **effort** do you make to assess results of your program?*
2. *Have you set **specific** objectives for your employee publication?*
3. *Is it **larger** or fancier than you really need?*

Complex Contexts.

1. *Have you permitted it to become a giveaway program rather than one that has the goal of improved employee morale and, consequently, **increased** productivity?*
2. *Can you consider restricting any additional employee benefits to those paid for by profit-sharing money, such as was **done** in the union contract recently signed by American Motors Corporation?*
3. *Youngish man on the make, Madden labeled him, and was ready to guess that in a correct, not too pushing fashion, the junior partner of the firm had **political** ambitions; that Mrs. Garth would be impeccably suitable as the wife of a rising young lawyer; that there were three children, two boys and a girl; that she was active in the Woman's Club and he in Lions, Rotary, and Jaycee; and finally, that neither of them had harbored an unorthodox opinion since their wedding day.*

Keyword-Replaced Contexts.

1. **Original:** *In 1610, Hudson was probably in his **early** forties, a good navigator, a stubborn voyager, but otherwise fatally unsuited to his chosen profession.*

   Keywords: Hudson, was, in, his, forties

   Rewritten: *Hudson was in his **early** forties.*

2. **Original:** *He alleged no **specific** procedural errors or evidence withheld; nor did he elaborate just what favorable evidence the Federal Bureau of Investigation reports might disclose.*

   Keywords: He, alleged, no, procedural, errors, withheld

   Rewritten: *He alleged no **specific** procedural errors withheld.*

3. **Original:** *Indeed, in some periods of our history and in some neighborhoods the job opportunities have been so good that undoubtedly a great many boys who were potential members of the professions quit school at an **early** age and went to work.*

   Keywords: boys, quit, school, at, an, age

   Rewritten: *boys quit school at an **early** age.*

Across the 150 instances used for replacement, the average proportion of keywords in the original contexts (*Density*) is 0.27, which increases to 1.0 after replacement.