GenDecider: Integrating "None of the Candidates" Judgments in Zero-Shot Entity Linking Re-ranking

Anonymous ACL submission

Abstract

We introduce GenDecider, a novel re-ranking approach for Zero-Shot Entity Linking (ZSEL), built on the Llama model. It innovatively detects scenarios where the correct entity is not among the retrieved candidates, a common oversight in existing re-ranking methods. By autoregressively generating outputs based on the context of the entity mention and the candidate entities, GenDecider significantly enhances disambiguation, improving the accuracy and reliability of ZSEL systems, as demonstrated on the benchmark ZESHEL dataset.

1 Introduction

Zero-Shot Entity Linking (ZSEL) (Logeswaran et al., 2019), a crucial task in NLP, links entity mentions in text to corresponding entities in a Knowledge Base (KB), when no labeled examples of those entities are available. The importance of this task stems from its ability to handle entities dynamically, particularly in evolving KBs where new entities frequently emerge.

The prevailing approach in ZSEL, exemplified by the BLINK method (Wu et al., 2020), adopts a two-step process: initial retrieval of candidate entities followed by a re-ranking phase. While extensive research has improved the retrieval stage (Ma et al., 2021; Agarwal et al., 2022; Sui et al., 2022; Sun et al., 2022; Wu et al., 2023), the re-ranking phase, which is critical for final decision-making, has received comparatively less attention.

Moreover, a significant oversight in existing re-ranking studies (Wu et al., 2020; Tang et al., 2021; Barba et al., 2022; Xu et al., 2023) is the assumption that the correct entity is always among the retrieved candidates. This assumption, however, often does not hold in zero-shot settings, leading to the prevalence of what we call "None of the Candidates" (**NoC** for short) cases. When the correct entity is not among the retrieved candidates, opting for a NoC prediction is more beneficial than forcibly making a false positive prediction in real applications. Having NoC predictions can also offer feedback to the retrieval phase by highlighting the limitations of retrievers in zero-shot settings.

This paper introduces GenDecider, a novel approach that integrates NoC judgments into the ZSEL re-ranking process. GenDecider formulates the re-ranking task as a generative process using the recent Llama model (Touvron et al., 2023). Given the context of an entity mention and the retrieved candidates, GenDecider autoregressively generates an output that is either the ID of the correct entity candidate or a NoC judgment. This approach allows for direct interactions between the mention context and the candidates within the same input, facilitating more accurate disambiguation. Moreover, by supporting NoC judgments, GenDecider enhances the reliability of ZSEL systems.

The contributions of this work are twofold. Firstly, it presents a novel re-ranking formulation that addresses a significant gap in existing research by effectively detecting NoC scenarios. Secondly, the proposed method demonstrates a comprehensive approach to disambiguation, improving both the accuracy and applicability of ZSEL systems.

2 Related Work

Entity Linking (EL) methods can be broadly classified into generation-based and retrieval-based. Generation-based methods, such as GENRE (De Cao et al., 2020), directly generate entity titles but struggle with new entities in zero-shot settings (Xu et al., 2023). In contrast, retrieval-based methods, more suitable for zero-shot settings, follow a two-step approach: candidate retrieval and re-ranking. We focus on the re-ranking phase.

ZSEL Re-ranking. The ZSEL task, initiated by Logeswaran et al. (2019), challenges EL systems' capability to link new, unseen entities using minimal information, typically just brief entity descriptions from KBs. Notable works in ZSEL re-ranking include BLINK (Wu et al., 2020) which employs a Cross-Encoder for comprehensive analysis between mention contexts and entity descriptions. Bi-MPR (Tang et al., 2021) utilizes a bidirectional multi-paragraph reading model for deeper semantic understanding, while ReS (Xu et al., 2023) focuses on enhancing cross-entity comparisons. These approaches typically re-rank using similarity scores. ExtEnD (Barba et al., 2022) offers an alternative by formulating re-ranking as a text extraction task, not relying on entity descriptions.

Difference from NIL. The concept of NIL in EL refers to instances where an entity mention does not correspond to any entity in the entire KB (Zhu et al., 2023). It signifies that the mention either refers to an entity not present in the KB or is not an entity. In contrast, NoC indicates that the correct entity does exist in the KB but was not included in the candidate set by the retrieval model. Therefore, while NIL concerns the absence of a corresponding entity in the KB, NoC deals with missing the correct entity in the retrieval process.

3 Methodology

3.1 Task Definition

EL associates detected entity mentions in text with corresponding entities in KBs, typically through a two-step process: retrieval and re-ranking.

Retrieval: This phase aims to identify a set of candidate entities $\mathcal{C}(m)$ from the KB $\mathcal{E}$ for an entity mention m in a document d.

Re-ranking: Following retrieval, this phase targets re-evaluating the candidate entities $\mathcal{C}(m)$ to accurately identify the correct entity e.

ZSEL is characterized by that the training and test datasets do not share entities, mirroring real-world scenarios where new, unseen entities frequently emerge. Formally, let $\mathcal{E}_{train}$ and $\mathcal{E}_{test}$ represent the training and test KBs, respectively, with $\mathcal{E}_{train} \cap \mathcal{E}_{test} = \emptyset$. Each entity e in either $\mathcal{E}_{train}$ or $\mathcal{E}_{test}$ is associated with a textual description $Desc(e)$. Let $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$ be the corresponding sets of training and test documents. The objective of ZSEL is to train a retrieval-reranking system using $\mathcal{D}_{train}$ and $\mathcal{E}_{train}$, and then apply it to $\mathcal{D}_{test}$ and $\mathcal{E}_{test}$.

3.2 Integrating NoC into ZSEL Re-ranking

In this paper, we focus on enhancing the re-ranking phase. Traditional re-ranking methods typically assume that the correct entity is always present within the retrieved candidate set, which leads to a forced selection from this set. However, this assumption often does not hold in ZSEL scenarios, where the retrieval model (trained on $\mathcal{E}_{train}$) is more likely to fail to include the correct entity in the candidate set from $\mathcal{E}_{test}$ compared to traditional EL. Consequently, this leads to a higher rate of false positives in the final linking predictions, thereby affecting the reliability of EL systems.

To tackle this challenge, we propose integrating the NoC option into re-ranking. We reformulate re-ranking as a generative task, employing a decoder-only architecture, which allows the model to directly reason over the mention context and candidate entities within the same input.

The input in our formulation includes a task-specific instruction $Inst$, the context of the entity mention $Ctxt(m)$, and the set of retrieved candidates $\mathcal{C}(m)$. The generated output is either the ID of the correct entity $e \in \mathcal{C}(m)$, or a "None" designation when the correct entity is not among the candidates. This is formally represented as:

$$f : (Inst, Ctxt(m), \mathcal{C}(m)) \rightarrow ID(e) \; or \; None,$$

where $(Inst, Ctxt(m), \mathcal{C}(m))$ collectively forms the prompt for our re-ranking process.

3.3 GenDecider

In our empirical investigations, we discovered that recent open-sourced, decoder-only large language models (LLMs) struggled with our re-ranking formulation through In-Context Learning (ICL). This shortfall is likely attributable to their pre-training regimes, which may not heavily focus on disambiguation tasks. To overcome this limitation, we opted to fine-tune such an advanced LLM using Low-Rank Adaptation (LoRA) (Hu et al., 2021). LoRA's adaptability allows us to retain the base model's capabilities while introducing a small, disambiguation-specialized adaptor, resulting in our new model, GenDecider.

For training GenDecider, we collect entity mentions along with their top k candidates (i.e., $|\mathcal{C}(m)| = k$) given by the retrieval step as the training set. The choice of k is constrained by the base model's maximum context length. Note that this training set includes instances where the correct entity e is not among $\mathcal{C}(m)$, leading to NoC scenarios. Instances serve to form the following prompt for GenDecider:

Entity Mention: m
Entity Mention Context: $Ctxt(m)$

Based on the above entity mention and its context, identify the ID of the candidate in the following to which the entity mention refers (if none of them, assign the ID as "None"):

ID: $ID(e_0)$
Entity: e_0
Entity Description: $Desc(e_0)$

(omit other $k-1$ entity candidates)

During training, we direct the model to generate a JSON object, for example, `{"ID": "123"}` or `{"ID": "None"}`, facilitating easy post-processing.

During inference, to improve the likelihood of including the correct entity in the candidate set, the number of candidates $|\mathcal{C}(m)|$ can be larger than k. Since GenDecider cannot process at once, we employ a block-wise approach. We split the candidates into $\lceil |\mathcal{C}(m)|/k \rceil$ blocks for block-wise inference. Each block yields either a candidate prediction or a NoC prediction. We merge non-NoC predictions into a new set $\mathcal{C}'(m)$. If $|\mathcal{C}'(m)|$ still exceeds k, the process is repeated until the set meets the size criteria ($\leq k$). A final inference is then conducted on this set to get the ultimate prediction.

4 Experiments

4.1 Datasets

We conduct experiments on the widely-used ZSEL dataset ZESHEL (Logeswaran et al., 2019). Statistics of this dataset are listed in Appendix 8.1. All the mentions have their correct entities in the KBs, which allows us to experiment with NoC scenarios.

4.2 Evaluation Metrics

Normalized accuracy is traditionally used in re-ranking evaluations, measuring the performance only on the **subset** of test instances for which the correct entity is within the retrieved candidates by the retrieval step. However, this metric falls short for our re-ranking methodology as it does not consider NoC cases, which are pivotal in our approach.

Instead, we adopt precision, recall, and the F1 score as our primary evaluation metrics on the **entire** test set. Precision can reflect if the NoC cases are accurately predicted. Recall measures whether the model accurately identifies the correct entity when it is within the candidate set. It's noteworthy that recall is essentially equivalent to normalized accuracy. The F1 score offers a balanced measure of the model's overall performance, measuring its ability to identify correct entities and recognize when none of the candidates are suitable.

4.3 Setups

We implemented GenDecider on Vicuna-7B-v1.5, based on Llama 2 (Touvron et al., 2023), with a limit of 4096 tokens. For fine-tuning, we utilized the FastChat package, which supports LoRA. We set LoRA parameters to $r = 8$ and $\alpha = 16$, resulting in an adaptor with 4 million trainable parameters. Diverging from baseline methods that train and test on the top 64 BM25-retrieved candidates, GenDecider uses the top 10 candidates (i.e., $k = 10$) from BM25 for training, while for testing, we align with the baselines by using the top 64 candidates. Both mention contexts and entity descriptions were limited to 256 tokens.

The ZESHEL training dataset consists of 49,275 examples, including 30,614 examples where the correct entity is among the top 10 candidates and 18,661 examples where it is not, as identified by BM25. The training was conducted over 2 epochs with a batch size of 1. Checkpoint selection was guided by loss convergence on a 2% held-out subset of our training data, differing from baseline methods that use the ZESHEL validation set. This selection is designed to better simulate a general zero-shot setting. All experiments were conducted on a single NVIDIA A100-SXM4-80GB GPU.

4.4 Baselines Considering NoC

As NoC is a novel aspect in re-ranking studies, there are no existing baselines explicitly designed for it. For BLINK and ReS, which use scoring for re-ranking, we can introduce a thresholding mechanism to determine NoC. For an entity mention, if scores for all candidates are lower than the threshold, this instance is considered NoC. We conducted a grid search for thresholds (ranging from 0.1 to 0.9) on a subset of 500 training examples, aiming to maximize the F1 score. The best thresholds identified were 0.9 for BLINK and 0.1 for ReS. Bi-MPR was not included due to the unavailability of its code. Additionally, we incorporate the base model Vicuna-7B-v1.5 through ICL (Vicuna-ICL) with a modified prompt from Section 3.3 by appending a suffix instruction: `Only output the ID in this format {"ID": ""}`, guiding its decoding.

4.5 Results

Performance on ZEHSEL Test Sets. The BM25 retrieval results (Recall@64) on ZESHEL test sets

Method	Forgotten Realms	Lego	Star Trek	YuGiOh	Macro Avg.
Not Considering NoC					
BLINK	72.33/ 86.80/ 78.91	62.05/ 76.39/ 68.48	51.36/ 77.95/ 61.92	41.05/ 67.46/ 51.04	56.70/ 77.15/ 65.36
Bi-MPR*	74.67/ 89.60/ 81.46	65.39/ 80.50/ 72.16	53.39/ 81.04/ 64.37	41.83/ 68.74/ 52.01	58.82/ 79.97/ 67.78
ExtEnD*	66.35/ 79.62/ 72.38	52.96/ 65.20/ 58.45	48.24/ 73.21/ 58.16	36.51/ 60.01/ 45.40	51.02/ 69.51/ 58.85
ReS*	73.42/ 88.10/ 80.09	63.72/ 78.44/ 70.32	53.82/ 81.69/ 64.89	46.15/ 75.84/ 57.38	59.28/ 81.02/ 68.47
Vicuna-ICL (w/o none)	35.03/ 42.00/ 38.20	22.45/ 27.62/ 24.77	23.64/ 35.80/ 28.47	13.30/ 21.78/ 16.52	23.61/ 31.80/ 27.10
GenDecider (w/o none)	75.98/ **91.10**/ 82.86	66.14/ **81.42**/ 72.99	54.50/ **82.48**/ 65.63	46.40/ **75.99**/ 57.62	60.76/ **82.75**/ 70.07
Considering NoC					
BLINK-Thresholding	**88.34**/ 80.30/ 84.13	**81.96**/ 66.22/ 73.25	71.02/ 71.35/ 71.18	62.97/ 59.47/ 61.17	**76.07**/ 69.33/ 72.54
ReS-Thresholding	85.95/ 77.10/ 81.28	78.20/ 61.50/ 68.85	73.97/ 70.09/ 71.98	62.21/ 68.87/ 65.37	75.12/ 69.39/ 72.12
Vicuna-ICL	35.10/ 37.80/ 36.40	22.09/ 24.54/ 23.25	22.46/ 30.92/ 26.02	14.09/ 21.44/ 17.01	23.44/ 28.67/ 25.79
GenDecider	86.26/ 86.00/ **86.13**	79.06/ 72.90/ **75.85**	**74.75**/ 79.61/ **77.10**	**63.60**/ 73.11/ **68.03**	75.92/ 77.91/ **76.90**

Table 1: Performance (Precision/ Recall/ F1) on ZESHEL test datasets. * means results reported in Xu et al. (2023).

can be found in Appendix 8.2, showing the prevalence of NoC cases. Table 1 offers a snapshot of the current state of ZSEL re-ranking methods. In the group of not considering NoC, "(w/o none)" implies removing the instruction highlighted in blue from the prompt in both training and testing. Vicuna-ICL (w/o none) underperforms, showing the base model's limitations in this disambiguation task. In contrast, GenDecider (w/o none) excels in this group, achieving the highest scores across datasets, underscoring the effectiveness of task-oriented fine-tuning and the advantages of larger language models in complex disambiguation tasks.

Introducing NoC predictions significantly improves precision, suggesting a reduction in false positives across most methods. This improved precision, coupled with robust recall rates, leads to notably higher F1 scores, demonstrating the importance of NoC in achieving a more balanced performance. GenDecider shines in the NoC-inclusive group, topping F1 scores, maintaining strong recall, and achieving high precision, which affirms its efficacy and reliability in practical EL tasks when NoC is common. These insights confirm the crucial role of both model architecture and fine-tuning for achieving accurate disambiguation.

Category-Specific Performance. Table 2 presents F1 scores on the ZESHEL test sets by mention-entity overlap categories, including High Overlap (HO), Multiple Categories (MC), Ambiguous Substring (AS), and Low Overlap (LO), where LO poses the greatest challenge and constitutes 59% of the ZESHEL dataset. Details can be found in Appendix 8.3. Here the LO category particularly benefits from the NoC consideration, with GenDecider achieving the highest F1 score, underscoring its efficacy in challenging disambiguation tasks.

Robustness across Retrieval Methods. Figure 1 demonstrates the stability of re-ranking methods when applied to different retrieval strategies on the ZESHEL test sets. Despite being trained with BM25-retrieved candidates, these methods exhibit consistent performance when assessed with Bi-Encoder-retrieved candidates (Wu et al., 2020), showcasing their capacity to handle diverse candidate sets. GenDecider, in particular, retains high F1 scores across both retrieval methods, reinforcing its effectiveness amidst varying retrieval situations.

Method	HO	MC	AS	LO
BLINK	93.96	69.23	73.21	51.77
Bi-MPR*	92.50	75.23	70.85	52.04
ReS*	**94.08**	74.64	71.25	53.90
GenDecider (w/o none)	91.06	75.71	78.30	55.31
BLINK-Thresholding	90.84	74.44	74.05	62.20
ReS-Thresholding	88.52	74.81	70.45	63.71
GenDecider	90.69	**78.59**	**79.42**	**68.95**

Table 2: Category-specific performance (F1).

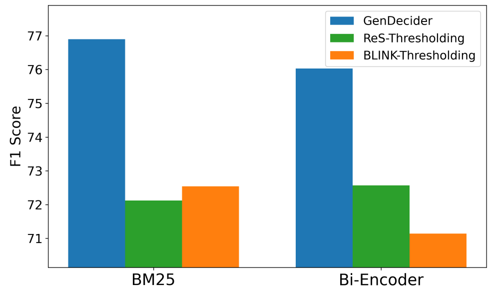

Figure 1: Robustness across retrieval methods.

5 Conclusion

This paper presents GenDecider, an innovative re-ranking approach for ZSEL that adeptly incorporates NoC judgments. Our extensive experiments on ZESHEL demonstrate that GenDecider achieves superior performance in challenging disambiguation scenarios. The results underscore the importance of NoC consideration in improving the reliability in the re-ranking phase.

6 Limitations

This study introduces GenDecider, a 7B-parameter model demonstrating state-of-the-art performance in zero-shot entity linking. However, there are limitations to consider.

Computational Efficiency: Due to its large size, GenDecider is computationally intensive, which may not be feasible for systems requiring real-time or online processing. Its deployment in environments with limited computational resources could be challenging, potentially limiting its practicality for certain applications.

Disambiguation Mechanism: While GenDecider shows promise, the underlying mechanisms of its disambiguation process may deserve further investigation. A deeper understanding of how GenDecider differentiates between entities could lead to improvements in both model efficiency and interpretability.

Future work should focus on enhancing the model's computational efficiency and exploring the disambiguation mechanism in more detail, which may yield more lightweight and interpretable models without compromising performance.

7 Ethics Statement

We comply with the ACL Code of Ethics.

References

Dhruv Agarwal, Rico Angell, Nicholas Monath, and Andrew McCallum. 2022. Entity linking via explicit mention-mention coreference modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4644–4658.

Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2022. Extend: Extractive entity disambiguation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2478–2488.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval. In *International Conference on Learning Representations*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3449–3460.

Xinyin Ma, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Weiming Lu. 2021. Muver: Improving first-stage entity retrieval with multi-view entity representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2617–2624.

Xuhui Sui, Ying Zhang, Kehui Song, Baohang Zhou, Guoqing Zhao, Xin Wei, and Xiaojie Yuan. 2022. Improving zero-shot entity linking candidate generation with ultra-fine entity type information. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2429–2437.

Kai Sun, Richong Zhang, Samuel Mensah, Yongyi Mao, and Xudong Liu. 2022. A transformational biencoder with in-domain negative sampling for zero-shot entity linking. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1449–1458.

Hongyin Tang, Xingwu Sun, Beihong Jin, and Fuzheng Zhang. 2021. A bidirectional multi-paragraph reading model for zero-shot entity linking. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13889–13897.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407.

Taiqiang Wu, Xingyu Bai, Weigang Guo, Weijie Liu, Siheng Li, and Yujiu Yang. 2023. Modeling fine-grained information via knowledge-aware hierarchical graph for zero-shot entity retrieval. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 1021–1029.

Zhenran Xu, Yulin Chen, Baotian Hu, and Min Zhang. 2023. A read-and-select framework for zero-shot entity linking. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13657–13666.

Fangwei Zhu, Jifan Yu, Hailong Jin, Juanzi Li, Lei Hou, and Zhifang Sui. 2023. Learn to not link: Exploring nil prediction in entity linking. *arXiv preprint arXiv:2305.15725*.

Method	High Overlap	Multiple Categories	Ambiguous Substring	Low Overlap
BLINK	93.62/ **94.30/ 93.96**	64.00/ 75.40/ 69.23	67.52/ 79.95/ 73.21	39.96/ 73.50/ 51.77
Bi-MPR*	92.17/ 92.84/ 92.50	69.54/ 81.93/ 75.23	65.34/ 77.37/ 70.85	40.17/ 73.88/ 52.04
ReS*	93.74/ 94.42/ 94.08	69.00/ 81.29/ 74.64	65.71/ 77.80/ 71.25	41.60/ 76.51/ 53.90
GenDecider (w/o none)	90.73/ 91.39/ 91.06	70.06/ **82.36**/ 75.71	73.68/ **83.53**/ 78.30	42.74/ **78.37**/ 55.31
BLINK-Thresholding	**95.58**/ 86.55/ 90.84	**80.18**/ 69.46/ 74.44	**81.61**/ 67.78/ 74.05	61.45/ 62.97/ 62.20
ReS-Thresholding	93.41/ 84.12/ 88.52	75.53/ 74.11/ 74.81	76.99/ 64.93/ 70.45	64.02/ 63.40/ 63.71
GenDecider	91.70/ 89.70/ 90.69	77.18/ 80.05/ **78.59**	81.09/ 77.80/ **79.42**	**66.51**/ 72.76/ **68.95**

Table 3: Category-specific performance (Precision/ Recall /F1) on ZESHEL test datasets. * means results reported in (Xu et al., 2023).

Domains	Entities	Mentions
Training		
American Football	31929	3898
Doctor Who	40281	8334
Fallout	16992	3286
Final Fantasy	14044	6041
Military	104520	13063
Pro Wrestling	10133	1392
StarWars	87056	11824
World of Warcraft	27677	1437
Validation		
Coronation Street	17809	1464
Muppets	21344	2028
Ice Hockey	28684	2233
Elder Scrolls	21712	4275
Test		
Forgotten Realms	15603	1200
Lego	10076	1199
Star Trek	34430	4227
YuGiOh	10031	3374

Table 4: Statistics of the ZESHEL dataset.

8 Appendix

8.1 ZESHEL Dataset

The statistics of the ZESHEL dataset (Logeswaran et al., 2019) are presented in Table 4. All the mentions have their correct entities in the corresponding KBs, which allows us to experiment with NoC scenarios.

8.2 Retrieval Results

Table 5 showcases the recall@64 performance for BM25 and Bi-Encoder on the ZESHEL test sets. The Bi-Encoder, as detailed in Wu et al. (2020), benefits from training on the ZESHEL training sets and its ability to capture semantics, yielding an enhanced retrieval efficacy over BM25. However, both methods struggle in Star Trek and YuGiOh, which indicates the prevalence of NoC cases.

Dataset	BM25	Bi-Encoder
Forgotten Realms	83.33	89.75
Lego	81.23	88.32
Star Trek	65.89	78.94
YuGiOh	60.85	65.65

Table 5: Retrieval performance (Recall@64) on ZESHEL test datasets.

8.3 Category-Specific Performance

Mentions in the ZESHEL dataset (Logeswaran et al., 2019) are categorized based on the token overlap with their corresponding entities:

High Overlap (HO): The entity title is identical to the mention text.

Multiple Categories (MC): The entity title consists of the mention text followed by a disambiguation phrase (e.g., for the mention 'Batman', the title is 'Batman (Lego)').

Ambiguous Substring (AS): The mention is a substring of the entity title (e.g., the mention 'Agent' corresponds to the title 'The Agent').

Low Overlap (LO): All other mentions that do not fit the above categories are considered low overlap.

These categories represent roughly 5%, 28%, 8%, and 59% of the dataset's mentions, respectively. Table 3 presents detailed performance evaluations of precision, recall, and F1 scores on the ZESHEL test sets by mention-entity overlap categories.