kevinyu66 at SemEval-2026 Task 3: A Retrieval-Augmented LLM System for Aspect–Opinion Triplet Extraction

Kuan-Lin Yu[1,†,*] **Wen-Ni Liu**[2,†]
[1]National Cheng Kung University , [2]National Yang Ming Chiao Tung University
[†]Equal contribution [*]Corresponding author
***Contact:** guanlinyu4@gmail.com

Abstract

This paper describes our system used in the SemEval-2026 Task 3: Dimensional Aspect-Based Sentiment Analysis. To address the inherent subjectivity and nuanced emotional expressions in this task, we propose a Retrieval-Augmented Generation (RAG) framework based on Large Language Models (LLMs) for sentiment triplet extraction. Our approach leverages a dynamic retrieval mechanism to identify semantically similar training examples, which are then integrated into the prompts as in-context demonstrations. This strategy effectively guides the model's inference process by providing relevant linguistic patterns and emotional contexts. Our implementation is available at https://github.com/Kevinyu66/dimaste.

1 Introduction

Aspect-Based Sentiment Analysis (ABSA) (Liu et al., 2020; Brauwers and Frasincar, 2022) has evolved from discrete polarity classification to Dimensional ABSA (Lee et al., 2026). Traditional ABSA methods assign categorical sentiment labels (e.g., positive, negative, neutral) to aspect terms, which fail to capture the full spectrum of human emotions. The dimensional approach addresses this limitation by representing affective states along continuous axes, notably Valence (the degree of pleasantness) and Arousal (the level of activation or intensity) (Russell, 1980), offering a more fine-grained perspective that captures subtle emotional fluctuations in naturalistic language (Buechel and Hahn, 2017; Lee et al., 2022).

Building on this paradigm, SemEval-2026 Task 3 introduces the Dimensional ABSA (DimABSA) track (Yu et al., 2026). Grounded in the circumplex model of affect, this task requires predicting real-valued Valence and Arousal (VA) scores across diverse domains including Restaurant, Laptop, and Hotel in multiple languages. Our work specifically targets Subtask 2, Dimensional Aspect Sentiment Triplet Extraction (DimASTE), which involves the joint extraction of (aspect term, opinion term, VA score) triplets.

While Large Language Models (LLMs) have demonstrated strong few-shot and instruction-following capabilities (Brown et al., 2020; OpenAI et al., 2024), effectively leveraging them for this specific multidimensional extraction remains difficult. Two critical obstacles persist. First, as noted by (Ruan et al., 2025), sentiment annotations are inherently subjective, leading to significant variance across datasets. Second, a static fine-tuned model often struggles to adapt to the diverse linguistic contexts of unseen test queries, particularly in a multilingual and multi-domain setting (Scaria et al., 2024).

To address these issues, we argue that fine-tuning a specialized model on domain-specific data is essential, but its performance can be further amplified through dynamic context. Inspired by the retrieval-augmented prompting paradigm (Jian et al., 2025), we propose a hybrid framework that bridges supervised fine-tuning (SFT) and In-Context Learning (ICL).

Experimental results demonstrate that this hybrid approach consistently outperforms standard fine-tuning, particularly in capturing the fine-grained nuances of dimensional sentiment where annotation consistency is critical.

2 System Overview

We propose a retrieval-augmented triplet extraction framework, as illustrated in Figure 1. The architecture is designed to bridge the gap between static model weights and the subjective nature of dimensional sentiment annotations. The proposed system comprises three primary stages: Semantic Index-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 108–114

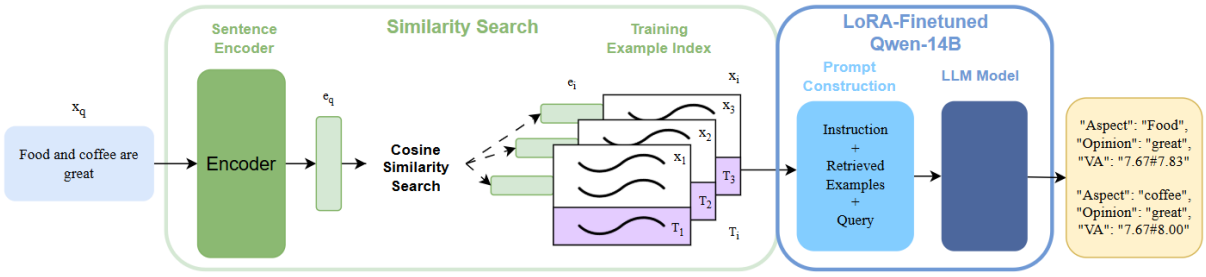

Figure 1: Architecture of our system.

ing, Instruction Tuning, and Retrieval-Augmented Inference.

2.1 Instruction Tuning

We select Qwen2.5-14B-Instruct (Yang et al., 2024) as our backbone model and perform supervised fine-tuning (SFT) using LoRA (Yu et al., 2023). The primary goal of this stage is to internalize the task logic and enforce a strict JSON output format.During training, we use a fixed prompt template in Table 1.

Given a training pair (x, T), where x denotes the input sentence and T the corresponding set of annotated aspect-opinion triplets. The model is optimized using the **cross-entropy** loss:

$$\mathcal{L} = -\sum_{t=1}^{|y|} \log p_\theta(y_t \mid y_{<t}, x)\,, \qquad (1)$$

where y represents the serialized JSON sequence of annotated triplets. By training exclusively on the core task format, the model establishes a "structured baseline" that we subsequently enhance with dynamic context during inference.

Prompt Template
System Prompt: Extract sentiment triplets from the input. Output exactly one JSON object. No explanations. If none, output {"Triplet": []}.
User Prompt: Extract (Aspect, Opinion, VA) explicitly mentioned. No inference. VA format: "valence#arousal" (1.00–9.00). Combine all results into ONE JSON.

Table 1: Prompt template for Instruction Tuning.

2.2 Semantic Indexing and Retrieval

To support retrieval-augmented inference, we construct a semantic index over the training set to enable the selection of semantically similar exemplars. This process ensures that the model is guided by in-domain demonstrations that align with the specific labeling style of the dataset, effectively mitigating the challenges of annotation subjectivity.

Algorithm 1: Retrieval-Augmented Triplet Extraction

Input: Query sentence x_q, Semantic index $\mathcal{I} = \{(e_i, x_i, T_i)\}_{i=1}^{N}$, Fine-tuned LLM $\mathcal{M}_\theta$
Output: Predicted triplets T_q

// Phase 1: Semantic Retrieval
$e_q \leftarrow \text{SentenceEncoder}(x_q)$
$S \leftarrow \{(e_q^\top e_i, x_i, T_i) \mid (e_i, x_i, T_i) \in \mathcal{I}\}$
$\mathcal{D}_{topK} \leftarrow \text{SelectTopK}(S, K)$
// Phase 2: Prompt Composition
$P \leftarrow$ Instruction
for $(x_j, T_j) \in \mathcal{D}_{topK}$ **do**
  $P \leftarrow P \oplus \text{FormatExample}(x_j, T_j)$
end
$P \leftarrow P \oplus x_q \oplus$ "Output:"
// Phase 3: Generation
$y_{json} \leftarrow \mathcal{M}_\theta(P)$
$T_q \leftarrow \text{ParseJSON}(y_{json})$
return T_q

Indexing Phase Given a training dataset $\mathcal{D} = \{(x_i, \mathcal{T}_i)\}_{i=1}^{N}$, where x_i denotes the input sentence and T_i denotes the corresponding set of annotated aspect–opinion triplets. We encode each training sentence into a dense semantic representation using a multilingual sentence embedding model, **paraphrase-multilingual-mpnet-base-v2**. All sentence embeddings are L2-normalized and stored in a serialized tensor-based index. This design enables efficient similarity computation via inner product, which is equivalent to cosine similarity under normalization.

Inference Phase As detailed in Algorithm 1, for a query sentence x_q, we first compute its dense embedding e_q using the same encoder. We then retrieve the top-K most similar training instances from the index $\mathcal{I}$ by calculating the semantic similarity scores:

$$\text{score}(x_q, x_i) = e_q^\top e_i. \tag{2}$$

These retrieved examples (x_i, T_i) serve as the dynamic context required for the subsequent prompting stage.

2.3 Retrieval-Augmented Prompting

The final step of our pipeline is the construction of a retrieval-augmented prompt. As shown in Phase 2 of Algorithm 1, we aggregate the task instruction, the retrieved exemplars, and the query sentence. Notably, the underlying structure of the system and user prompts remains identical to those used during the instruction tuning phase (see Table 1) to ensure the model operates within its learned task distribution. The key distinction lies in the inclusion of dynamic context: we append K semantically similar demonstrations between the instructions and the target query. Formally, the prompt is defined as:

$$\text{Prompt}(x_q) = \text{Instr.} \parallel \bigoplus_{i=1}^{K} \text{Format}(x_i, T_i) \parallel x_q \tag{3}$$

where $\parallel$ denotes string concatenation and $\text{Format}(\cdot)$ transforms the retrieved triplet into the JSON demonstration style.

Table 2 provides a concrete example of this few-shot composition. In this example, when presented with a query regarding "screen brightness," the system retrieves a semantically related instance concerning "battery life." By observing the triplet structure and the specific valence-arousal (VA) scoring format (e.g., `4.17#5.83`) in the demonstration, the model is provided with a "subjective anchor" to calibrate its extraction behavior for the current query.

3 Experiments

3.1 Experimental Setup

Model Backbone. We employ Qwen2.5-14B-Instruct[1] (Yang et al., 2024) as our base model. We specifically selected it among models under 15B parameters, as it is currently the most downloaded model in its class on Hugging Face and demonstrated the best multilingual performance in our preliminary tests.

Training Configuration. To enable 14B model training on a single GPU, we use 4-bit quantization (NF4). We apply LoRA (Yu et al., 2023) to all attention projection layers with $r = 16$ and $\alpha = 32$. Training is conducted for 3 epochs using the AdamW optimizer with a learning rate of 2×10^{-4} and an effective batch size of 8. The model is optimized to generate structured JSON triplets under a causal language modeling objective.

Retrieval & Decoding. During inference, we use paraphrase-multilingual-mpnet-base-v2 [2] to retrieve the top-$K = 10$ training exemplars based on cosine similarity. To ensure output stability, we employ greedy decoding (temp=0). Outputs failing JSON parsing are treated as empty predictions.

All experiments are conducted on a single NVIDIA A100 (40GB) GPU.

3.2 Metrics

We evaluate our system using the Continuous F1 (cF1) score, the official metric for Subtask 2. Unlike standard F1, cF1 incorporates the prediction error of Valence-Arousal (VA) intensities. A prediction is a Continuous True Positive (cTP) only if its categorical elements (Aspect and Opinion) match the gold annotation:

$$cTP(t) = \begin{cases} 1 - \text{dist}(VA_p, VA_g), & t \in P_{cat} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

The term $\text{dist}(\cdot)$ represents the Euclidean distance between predicted (VA_p) and gold (VA_g) values, normalized by the maximum possible distance $D_{max} = \sqrt{128}$. Continuous Precision and Recall are computed by summing cTP over the prediction and ground-truth sets, respectively, with cF1 being their harmonic mean.

3.3 Result

The development set results in Table 3 are computed locally using the official evaluation script, while the test set results in Table 4 are obtained from the SemEval evaluation server, which additionally provides leaderboard rankings. Therefore, ranking information is only available for the test set.

[1] https://huggingface.co/Qwen/Qwen2.5-14B-Instruct

[2] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

Inference Prompt Example with ($K = 3$)	
Current Query	**Input:** The screen is surprisingly bright even in daylight. **Output:** (to be predicted)
Retrieved 1	**Input:** The battery life is fairly low. **Output:** `{"Triplet": [{"Aspect": "battery life", "Opinion": "fairly low", "VA": "4.17#5.83"}]}`
Retrieved 2	**Input:** The charging speed is quite impressive. **Output:** `{"Triplet": [{"Aspect": "charging speed", "Opinion": "impressive", "VA": "7.20#6.50"}]}`
Retrieved 3	**Input:** The display is a bit dim under direct sunlight. **Output:** `{"Triplet": [{"Aspect": "display", "Opinion": "dim", "VA": "3.10#4.20"}]}`

Table 2: An example of a retrieval-augmented inference prompt with K=3 retrieved demonstrations.

Language	Domain	cF1	cPrecision	cRecall	cTP
English	Laptop	0.6088	0.6004	0.6175	195.73
	Restaurant	0.7211	0.7141	0.7281	297.08
Japanese	Hotel	0.5344	0.5492	0.5205	189.46
Russian	Restaurant	0.4854	0.4561	0.5187	52.91
Chinese	Laptop	0.5009	0.4985	0.5034	255.21
	Restaurant	0.6110	0.6047	0.6174	469.84
Ukrainian	Restaurant	0.5025	0.4842	0.5222	53.27
Tatar	Restaurant	0.4436	0.4333	0.4545	46.36

Table 3: Scores on the development set (locally evaluated; ranking not available).

Table 3 and Table 4 present the overall performance comparison across different languages and domains. Our system consistently outperforms the provided baseline on all evaluated settings, demonstrating strong generalization ability across multilingual and multi-domain scenarios. In particular, substantial improvements are observed on low-resource languages such as Japanese, Russian, Tatar, and Ukrainian, indicating the effectiveness of our approach under limited supervision.

Although our system does not achieve the best performance in all settings, the performance gap between our method and the top-ranked system remains relatively small across most language–domain pairs. These results suggest that the proposed retrieval-augmented and fine-tuned LLM framework provides a competitive and robust solution for aspect–opinion triplet extraction. In the following section, we further analyze the contributions of LoRA fine-tuning and retrieval size through detailed ablation studies.

3.4 Ablation Study

We conduct an ablation study to evaluate how the number of retrieved demonstrations K influences the performance of our fine-tuned Qwen2.5-14B model. As shown in Table 5, incorporating semantically similar examples consistently improves the cF1 score compared to the zero-shot baseline ($K = 0$).The performance gains when moving from $K = 0$ to $K = 3$, where the model begins to leverage the retrieved "subjective anchors" for better VA intensity calibration. Increasing the retrieval size further to $K = 10$ yields the best overall performance, albeit with diminishing marginal returns. This suggests that while more contextual examples provide richer references for capturing annotation nuances, the model's inherent task-specific knowledge (acquired during LoRA fine-tuning) already provides a robust foundation, which the retrieval mechanism then successfully refines.

As shown in Table 6, opinion span errors are particularly frequent, which may be partly attributed to the strict exact-match evaluation protocol. Even

Language	Domain	cF1	cPrecision	cRecall	cTP	Rank
English	Laptop	0.5503	0.6176	0.4962	979.58	9 / 20
	Restaurant	0.6707	0.7353	0.6165	1312.53	**5 / 20**
Japanese	Hotel	0.5366	0.5497	0.5242	756.37	**6 / 17**
Russian	Restaurant	0.5117	0.5114	0.5121	670.90	9 / 17
Chinese	Laptop	0.4802	0.4800	0.4805	924.96	**6 / 15**
	Restaurant	0.5089	0.5084	0.5093	1457.17	7 / 15
Ukrainian	Restaurant	0.4865	0.5010	0.4727	619.27	9 / 15
Tatar	Restaurant	0.3731	0.3832	0.3636	476.31	10 / 16

Table 4: Official test set performance and leaderboard ranking (evaluated by the SemEval server).

Model	$K = 0$	$K = 3$	$K = 10$
Qwen2.5-14B	0.4832	0.4982	**0.5009**

Table 5: Ablation study on the impact of retrieval size K on the zho laptop development set for finetuned SLM.

Error Type	Count	Ratio (%)
VA Miscalibration	70	22.08
Opinion Span Error	65	20.50
Aspect Error	34	10.73
Extra Triplet	23	7.26
Missing Triplet	17	5.36

Table 6: Error distribution on the development set.

minor lexical variations (e.g., "bad enough" vs. "bad") are treated as incorrect, despite conveying similar sentiment semantics. This suggests that the evaluation may underestimate the model's ability to capture sentiment meaning when minor wording differences occur.

We also observe from Table 6 that extra triplets occur in a non-negligible proportion of cases. This aligns with the observation that the model's extraction behavior changes after fine-tuning: before fine-tuning, the model tends to produce fewer triplets, adopting a conservative strategy, whereas after LoRA fine-tuning, it becomes more recall-oriented, generating more aspect–opinion pairs but occasionally introducing extra predictions.

4 Related Work

Traditional ABSA has evolved from pipeline models to unified generative frameworks (Hu et al., 2022; Gou et al., 2023), but these methods mainly focus on categorical sentiment labels. In contrast, dimensional sentiment analysis models affective states on continuous Valence–Arousal (VA) scales (Russell, 1980; Buechel and Hahn, 2017; Lee et al., 2022, 2026), enabling finer-grained emotional representation.

Recent work adapts Large Language Models (LLMs) to sentiment tasks using parameter-efficient tuning methods such as LoRA (Ren and Sutherland, 2025; Zhuang et al., 2023). We follow this direction by applying LoRA to a 14B backbone to balance performance and efficiency in multilingual settings.

Retrieval-augmented prompting dynamically selects in-context examples to guide inference. While prior work (Jian et al., 2025) focuses on discrete classification, our approach employs lightweight semantic retrieval for structured triplet extraction, serving as a subjective anchor to better calibrate VA predictions under annotation variability (Ruan et al., 2025).

5 Conclusion

In this paper, we present a hybrid framework combining parameter-efficient fine-tuning and retrieval-augmented prompting for the DimASTE task. By formulating triplet extraction as structured generation, Qwen2.5-14B effectively learns complex task patterns through LoRA.

Our results show that while fine-tuning provides a strong baseline, semantic retrieval further improves overall performance. Retrieved examples provide contextual guidance, helping the model capture both structural consistency and sentiment intensity under annotation variability.

This work demonstrates that combining lightweight adaptation with dynamic retrieval is effective for dimensional sentiment analysis in multilingual and multi-domain settings.

References

Gianni Brauwers and Flavius Frasincar. 2022. A survey on aspect-based sentiment classification. *ACM Comput. Surv.*, 55(4).

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Sven Buechel and Udo Hahn. 2017. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 578–585.

Zhibin Gou, Qingyan Guo, and Yujiu Yang. 2023. MvP: Multi-view prompting improves aspect sentiment tuple prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4380–4397, Toronto, Canada. Association for Computational Linguistics.

Mengting Hu, Yike Wu, Hang Gao, Yinhao Bai, and Shiwan Zhao. 2022. Improving aspect sentiment quad prediction via template-order data augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7900, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhongquan Jian, Yanhao Chen, Jiajian Li, Shaopan Wang, Xiangjian Zeng, Junfeng Yao, Xinying An, and Qingqiang Wu. 2025. Simrp: syntactic and semantic similarity retrieval prompting enhances aspect sentiment quad prediction. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Lung-Hao Lee, Jian-Hong Li, and Liang-Chih Yu. 2022. Chinese emobank: Building valence-arousal resources for dimensional sentiment analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 21(4):65.

Lung-Hao Lee, Liang-Chih Yu, Natalia Loukashevich, Ilseyar Alimova, Alexander Panchenko, Tzu-Mi Lin, Zhe-Yu Xu, Jian-Yu Zhou, Guangmin Zheng, Jin Wang, Sharanya Awasthi, Jonas Becker, Jan Philip Wahle, Terry Ruas, Shamsuddeen Hassan Muhammad, and Saif M. Mohammad. 2026. Dimabsa: Building multilingual and multidomain datasets for dimensional aspect-based sentiment analysis. *Preprint*, arXiv:2601.23022.

Haoyue Liu, Ishani Chatterjee, MengChu Zhou, Xiaoyu Sean Lu, and Abdullah Abusorrah. 2020. Aspect-based sentiment analysis: A survey of deep learning methods. *IEEE Transactions on Computational Social Systems*, 7(6):1358–1375.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Yi Ren and Danica J. Sutherland. 2025. Learning dynamics of LLM finetuning. In *The Thirteenth International Conference on Learning Representations*.

Zhihao Ruan, Runyang You, Kaifeng Yang, Junxin Lin, Wenwen Dai, Mengyuan Zhou, Meizhi Jin, and Xinyue Mei. 2025. PAI at SemEval-2025 task 11: A large language model ensemble strategy for text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1136–1142, Vienna, Austria. Association for Computational Linguistics.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Kevin Scaria, Himanshu Gupta, Siddharth Goyal, Saurabh Sawant, Swaroop Mishra, and Chitta Baral. 2024. InstructABSA: Instruction learning for aspect based sentiment analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 720–736, Mexico City, Mexico. Association for Computational Linguistics.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

Liang-Chih Yu, Jonas Becker, Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Lung-Hao Lee, Ying-Lung Lin, Jin Wang, Jan Philip Wahle, Terry Ruas, Alexander Panchenko, Ilseyar Alimova, Kai-Wei Chang, Lilian Wanzare, Nelson Odhiambo, Bela Gipp, and Saif M. Mohammad. 2026. SemEval-2026 task 3: Dimensional aspect-based sentiment analysis (DimABSA). In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics.

Yu Yu, Chao-Han Huck Yang, Jari Kolehmainen, Prashanth G. Shivakumar, Yile Gu, Sungho

Ryu Roger Ren, Qi Luo, Aditya Gourav, I-Fan Chen, Yi-Chieh Liu, Tuan Dinh, Ankur Gandhe Denis Filimonov, Shalini Ghosh, Andreas Stolcke, Ariya Rastow, and Ivan Bulyko. 2023. Low-rank adaptation of large language model rescoring for parameter-efficient speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Shengyao Zhuang, Bing Liu, Bevan Koopman, and Guido Zuccon. 2023. Open-source large language models are strong zero-shot query likelihood models for document ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8807–8817, Singapore. Association for Computational Linguistics.