BIRCO: A Benchmark of Information Retrieval Tasks with Complex Objectives

Anonymous ACL submission

Abstract

We present the **B**enchmark of **I**nformation **R**etrieval (IR) tasks with **C**omplex **O**bjectives (BIRCO) to evaluate the ability of IR models to follow multi-faceted task objectives. We study the performance of various embedding, distilled and fine-tuned IR models on BIRCO, and find them lacking. We provide a unified framework for investigating the performance of large language models (LLMs) on these tasks. The proposed framework consists of 3 modular components: task-objective awareness; chain-of-thought reasoning; and task decomposition. We investigate the effects of these factors on LLM performance, and identify a simple baseline model which matches or outperforms existing approaches and more complex alternatives. No approach achieves satisfactory performance on all benchmark tasks, suggesting that stronger models and new retrieval protocols are necessary to address complex user needs. [1]

1 Introduction

Information retrieval (IR) tasks have traditionally been centered around matching queries with semantically similar passages. However, user objectives may go significantly beyond retrieving based on similarity. As a motivating example, consider a user who wants to find papers that refute a particular scientific claim. This would not be well-captured by similarity-driven search, which would also retrieve papers that support the claim. In addition, the user may have multiple objectives in their search. They may be searching for papers that measure the response of a drug in a specific population, using a certain set of measurements.

We propose the BIRCO benchmark for evaluating the performance of IR systems on tasks with *complex objectives*. We curate 5 open-source datasets (DORIS-MAE (Wang et al., 2023), ArguAna (Wachsmuth et al., 2018), WhatsThatBook (Lin et al., 2023), Clinical-Trial (Koopman and Zuccon, 2016), and RELIC (Thai et al., 2022)), which contain paragraph-length queries with multi-faceted task objectives. This represents a challenging test bed for methods that aim to address complex user search needs.

IR systems have branched into three primary categories: pre-trained embedding models, language models (encoder-decoder and decoder-only) that have been fine-tuned for IR tasks, and task-agnostic models based on Large Language Models (LLMs) like GPT-4. Our research aims to examine the performance of these models on the BIRCO benchmark. We focus on state-of-the-art models from all categories, investigating their ability to handle the complex requirements of BIRCO.

In order to study the factors that affect LLM performance on these tasks, we introduce a framework for constructing retrieval protocols. This framework varies four factors: task-specific descriptions of the search objective, ranking vs. direct scoring, chain-of-thought reasoning (Wei et al., 2022), and decomposition of complex tasks into subtasks. Through this study, we aim to provide a foundation for advancing IR systems, particularly in the realm of complex search objectives.

2 Related Work

IR Benchmarks

IR benchmarks such as MS MARCO (Nguyen et al., 2016), NQ (Kwiatkowski et al., 2019), LOTTE (Santhanam et al., 2022), BEIR (Thakur et al., 2021) and BERRI (Asai et al., 2023) consist mostly of sentence-level queries, and their task objectives, while varying to some degree, focus on finding semantically similar passages, with one exception: the ArguAna dataset (Wachsmuth et al., 2018), which is a counterargument retrieval task.

Complex Query IR Tasks

Several recent datasets (DORIS-MAE, WTB, Clinical-Trial, RELIC) pose more complex re-

[1] https://github.com/BIRCO-benchmark/BIRCO.git

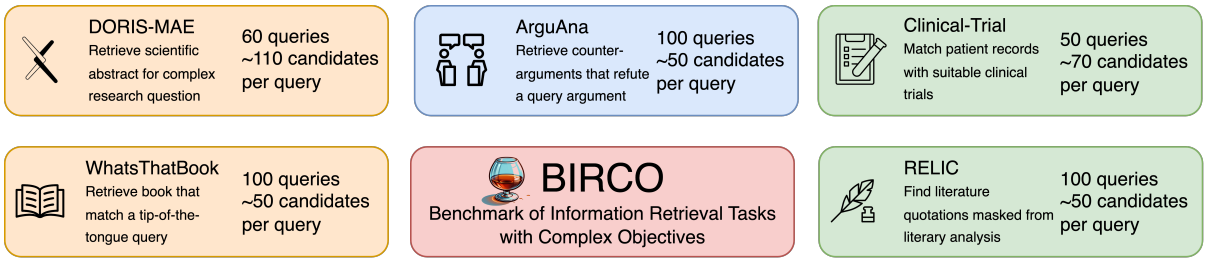

Figure 1: BIRCO contains 5 IR tasks with complex objectives

Models	MS MARCO MRR@10	BIRCO MRR@10
$\text{ANCE}_{\text{FirstP}}$	33.0	20.0
SimLM	41.1	18.1
SPLADE-v2	36.8	17.7

Models	NQ R@20	BIRCO R@20
$\text{ANCE}_{\text{FirstP}}$	81.9	49.6
SimLM	85.2	44.6
BM25	59.1	33.5

Models	BEIR R@10	BIRCO R@10
E5-L-v2	50.0	38.52
RankLLaMA	56.6	47.42
TART	44.8	39.48

Table 1: Comparing BIRCO's difficulty with other IR datasets. Models and metrics are chosen based on availability of data in the published literature.

trieval tasks (Wang et al., 2023; Lin et al., 2023; Koopman and Zuccon, 2016; Thai et al., 2022). In these datasets, the queries are paragraph-length, and passages should match the queries along multiple dimensions. See Figure 1.

Specialized retrieval models

Pretrained (Greene et al., 2022; Wang et al., 2022; Gao et al., 2021) and fine-tuned (Liu et al., 2023; Chuang et al., 2022; Dai et al., 2022; Gao et al., 2023; Ferraretto et al., 2023; Pan et al., 2023) embedding models have formed the core of most IR systems due to their speed and simplicity. More recently, there have been methods for fine-tuning language models for ranking and retrieval, including monoT5 (Nogueira et al., 2020) and Rank-Llama (Ma et al., 2023a). TART (Asai et al., 2023) and INSTRUCTOR (Su et al., 2023) are trained to follow task-specific instructions during retrieval.

LLM-based IR systems

Recent research has shown that LLMs can be effectively used for the re-ranking stage of IR. Sachan et al. (2022); Liang et al. (2022) compute a relevance score with output logits, Qin et al. (2023) use pairwise comparison among passages with open-source LLMs, Sun et al. (2023); Ma et al. (2023b) use list-wise comparisons, and Zhuang et al. (2023) have the LLM assign a 4-way label to each query/passage pair. These methods have primarily been evaluated on sentence-level queries.

3 Benchmark Construction

BIRCO is constructed to allow for statistically valid evaluation of model performance. Four of the five benchmark datasets do not have previously defined development set/test set splits. We therefore define splits for these datasets, ensuring that there is no overlap between queries or passages across the splits.

BIRCO is also designed specifically for benchmarking LLM performance. There are two distinctive issues that arise in this context. First, retrieval performance can be inflated due to data contamination from pretraining. In order to address this, we remove queries that GPT-4 can answer without access to passages. (This is most relevant for WTB and RELIC.) Second, it is prohibitively expensive to evaluate LLMs on the entire set of passages for each query. We therefore define candidate pools for each query, restricting LLM search to these smaller pools. This is standard for many other IR tasks, where it is known as the passage re-ranking stage.

3.1 Candidate Pool Construction

To make BIRCO more tractable for LLM-based retrieval, we construct candidate pools for each query using the lexicon-matching algorithm BM25 (Trotman et al., 2014) and state-of-the-art text embedding model ada-002 (Greene et al., 2022). Each query has a candidate pool of 50 passages. (The ground-truth passage is inserted when necessary.) As shown in Appendix Table 4, the difficulty is comparable to the original datasets for four of five tasks, and remains challenging for the fifth.

4 A Framework for LLM Re-Ranking

We investigate the effect of several factors on LLM retrieval performance.

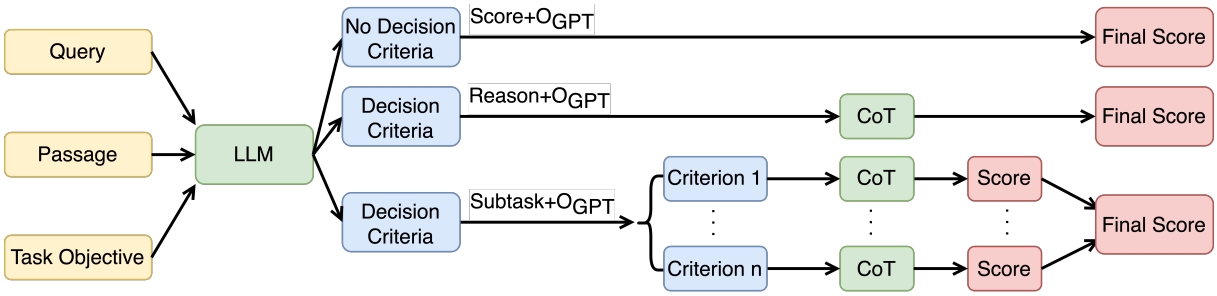

Figure 2: A framework for integrating LLMs into retrieval tasks

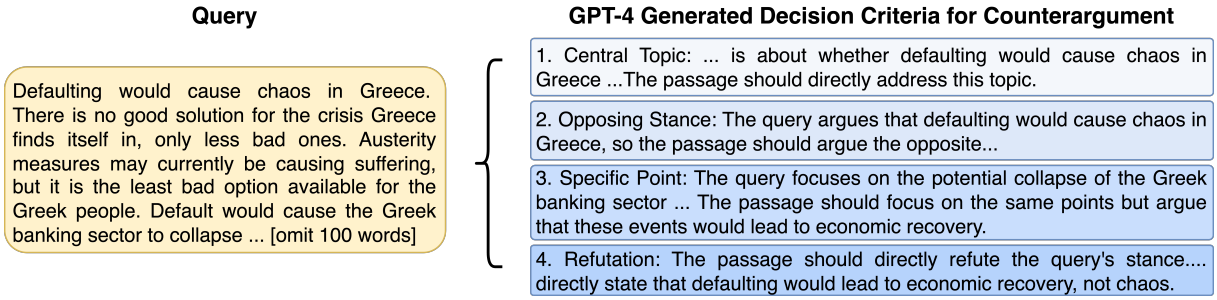

Figure 3: An example query and LLM-generated decision criteria

4.1 Task Objective Awareness

The tasks in BIRCO vary in their objectives, which can be clearly articulated in prompts to LLMs. Any model that uses a prompt containing the task description is suffixed with "O". Alternatively, as a simpler baseline, LLMs can be prompted to retrieve semantically similar passages without knowing the task objective. Prompts used for task objectives can be found in Appendix C.

4.2 Ranking vs. Scoring

LLMs can find relevant passages by either ranking them comparatively or by directly scoring them. Rank_{GPT} (Sun et al., 2023) puts a list of passages in the LLM prompt, and iteratively filters for the most relevant passages using a sliding window approach. LLMs can also score passages one at a time. For this direct scoring approach, the LLM is prompted to generate a numerical score from 0-10.

When passages are directly scored by the LLM, there can be ties in which several passages receive the same score. By default, we use E5-v2, an embedding model, for tie-breaking.

4.3 Explicit Reasoning

To investigate the role of natural language reasoning in complex query retrieval, we compare two approaches to scoring a query/passage pair. As shown in Figure 2, the first approach (shown in the middle of Figure 2, $\text{Reason+O}_{\text{GPT}}$) is to generate a set of decision criteria for judging whether a query is relevant to a passage. The LLM is instructed to follow its own decision criteria step-by-step before producing a final score. Another approach (top of Figure 2, $\text{Score+O}_{\text{GPT}}$) is to directly produce a score given the query, passage, and task objective. The detailed prompt structure for these two methods is recorded in Appendix D.

4.4 Task Decomposition

We investigate the effect of task decomposition, in which the LLM-generated decision criteria are used to define substasks which are independently solved by the LLM. The final score is computed by averaging scores from the subtasks. This strategy, denoted as $\text{Subtask+O}_{\text{GPT}}$, aims to reduce the complexity of evaluating whether a passage is relevant to a query.

5 Experiments

We use the GPT4-06-13 checkpoint as the LLM via OpenAI's API. See Appendix A for details about the baseline models, compute requirements, and additional embedding model experiments.

5.1 Results

Table 1 compares model performance on BIRCO to other IR benchmarks. Models perform significantly worse on BIRCO compared to published results on MS MARCO (Nguyen et al., 2016), NQ (Kwiatkowski et al., 2019), and BEIR (Thakur et al., 2021), validating the difficulty of BIRCO.

Model	DORIS-MAE		ArguAna		WTB		Clinical-Trial		RELIC	
	nDCG@10	R@5	nDCG@10	R@5	nDCG@10	R@5	nDCG@10	R@5	nDCG@10	R@5
Embedding Models										
E5-L-v2	72.0 ± 1.2	14.6 ± 1.9	43.5 ± 3.2	62.0 ± 4.8	36.6 ± 4.0	39.9 ± 4.8	29.4 ± 2.7	10.5 ± 1.7	11.1 ± 2.6	14.9 ± 3.4
SIMCSE	70.8 ± 1.3	14.4 ± 1.8	39.8 ± 3.5	51.1 ± 5.1	31.7 ± 4.0	37.0 ± 5.0	27.7 ± 2.7	10.5 ± 1.6	12.9 ± 2.5	15.6 ± 3.7
Promptagator	76.0 ± 1.4	15.6 ± 1.8	63.0 ± 3.3	**77.8** ± 4.2	59.1 ± 4.2	62.9 ± 4.7	NA	NA	NA	NA
Encoder-Decoder Models										
TART	64.9 ± 1.5	10.2 ± 1.7	47.6 ± 3.1	63.7 ± 4.8	42.7 ± 3.7	46.1 ± 4.9	32.8 ± 2.5	10.8 ± 1.3	9.4 ± 1.9	7.9 ± 2.7
TART+O	49.6 ± 1.8	4.9 ± 1.0	23.1 ± 2.8	28.0 ± 4.4	17.9 ± 2.4	20.7 ± 4.1	23.6 ± 2.2	7.7 ± 1.0	14.3 ± 2.3	18.3 ± 3.8
MonoT5	66.9 ± 1.5	12.4 ± 1.8	45.9 ± 3.2	52.0 ± 4.9	50.8 ± 4.2	59.0 ± 4.9	33.2 ± 2.5	**14.2** ± 2.3	13.9 ± 2.7	11.9 ± 3.2
Decoder-Only Model										
Rank-LLaMA	75.1 ± 1.3	**18.4** ± 2.0	53.2 ± 2.9	71.9 ± 4.4	63.7 ± 3.9	69.8 ± 4.6	29.7 ± 2.2	9.8 ± 1.1	15.4 ± 2.4	19.0 ± 4.0
Comparison-based LLM IR Systems										
Rank_{GPT}	76.2 ± 1.3	**17.5** ± 2.3	26.7 ± 2.2	17.0 ± 3.8	**79.5** ± 3.5	**85.9** ± 3.4	39.9 ± 2.6	**14.8** ± 1.4	40.1 ± 4.0	48.0 ± 5.1
$\text{Rank+O}_{\text{GPT}}$	77.4 ± 1.3	**17.6** ± 2.1	54.4 ± 3.2	70.9 ± 4.5	**82.1** ± 3.3	**88.0** ± 3.2	38.6 ± 2.8	14.6 ± 1.5	**62.3** ± 3.9	**70.1** ± 4.5
Scoring-based LLM IR Systems										
$\text{Score+O}_{\text{GPT}}$	**79.9** ± 1.2	**19.3** ± 2.0	51.6 ± 2.7	70.0 ± 4.6	**83.3** ± 3.1	**90.9** ± 2.8	**43.4** ± 2.4	**17.2** ± 1.6	**54.1** ± 3.3	**70.1** ± 4.6
$\text{Reason+O}_{\text{GPT}}$	74.9 ± 1.3	**18.1** ± 2.3	59.9 ± 2.9	**76.8** ± 4.2	74.9 ± 3.6	82.8 ± 3.7	**45.7** ± 3.1	**17.0** ± 1.7	39.9 ± 3.3	51.0 ± 5.0
$\text{Subtask+O}_{\text{GPT}}$	**78.5** ± 1.2	**18.5** ± 1.9	**69.5** ± 3.0	**85.9** ± 3.5	**79.6** ± 3.5	**85.9** ± 3.3	**43.5** ± 2.9	**16.5** ± 1.6	53.2 ± 3.5	**62.9** ± 4.8

Table 2: nDCG@10 and Recall@5 for the benchmark datasets. Bold indicates $p > 0.05$ compared to the highest numerical value indicated in red. Error bars are standard errors. The notation +O indicates task objective awareness.

Table 2 shows the results for the strongest embedding models and language models on BIRCO. Most GPT4-based IR strategies significantly outperform embedding or small (<10B) language models. Among LLM models, Rank_{GPT} performs most poorly, achieving notably weaker results on ArguAna and RELIC. This is the only model without task objective awareness for which we report results; other models without task objective awareness performed very poorly on the development sets, so they were excluded from further analyses for cost reasons.

$\text{Score+O}_{\text{GPT}}$ performs well on 4 out of 5 tasks. This is one of the simplest LLM models: its prompt describes the task objective, and the model directly outputs a score given a query, passage pair, without performing any reasoning.

The Subtask model, which decomposes the query into subtasks and evaluates each subtask separately, has strong performance on all datasets. However, it only exceeds the performance of other LLM models on ArguAna.

No model achieves strong performance on DORIS-MAE (as measured by recall) or Clinical-Trial.

6 Conclusion

We have introduced BIRCO, a benchmark for IR tasks with complex objectives. BIRCO includes scientific, medical, literary, and current-events retrieval tasks, and is significantly more challenging than previous IR benchmarks.

We found that embedding methods and small (<10B parameters) language models have weak performance on the BIRCO tasks. Methods that use LLMs for ranking have stronger performance, though none achieve strong results across all tasks.

We evaluated several hypotheses regarding LLM performance. First, providing clear instructions to the LLMs regarding task objectives was critical to achieving good performance. Second, we did not find evidence that ranking by comparing passages improved performance relative to directly scoring passages. Third, in contrast to results in many other NLP domains (Kojima et al., 2022; Huang et al., 2022), we did not find evidence that chain-of-thought reasoning improves retrieval performance. Finally, decomposing queries into subtasks improved performance on only a single dataset.

The results underscore the need to develop IR methods that go beyond similarity-based retrieval. Strong performance on BIRCO requires models that can understand multi-faceted user intents. While GPT-4-based methods had the strongest performance, even they did not achieve adequate performance across tasks. Furthermore, it is currently prohibitively expensive to perform inference with LLMs for all but the smallest IR tasks. The challenge of complex user objectives will require improvements in model abilities and efficiency.

7 Limitations

There are few existing datasets with complex queries or non-standard search goals. We hope our work can encourage more research and task creation in this area, increasing the number of benchmarked IR tasks with complex objectives.

Furthermore, LLM methods are computationally expensive and can only be effectively employed in the passage re-ranking stage of a multi-stage IR process.

8 Ethical Considerations

No ethical concerns for this work.

References

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. 2023. Task-aware retrieval with instructions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3650–3675, Toronto, Canada. Association for Computational Linguistics.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. DiffCSE: Difference-based contrastive learning for sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218, Seattle, United States. Association for Computational Linguistics.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples. In *The Eleventh International Conference on Learning Representations*.

Fernando Ferraretto, Thiago Laitz, Roberto Lotufo, and Rodrigo Nogueira. 2023. Exaranker: Synthetic explanations improve neural rankers. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2409–2414.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade v2: Sparse lexical and expansion model for information retrieval. *arXiv preprint arXiv:2109.10086*.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2022. From distillation to hard negative sampling: Making sparse neural ir models more effective. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2353–2359.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ryan Greene, Ted Sanders, Lilian Weng, and Arvind Neelakantan. 2022. New and improved

embedding model. https://openai.com/blog/new-and-improved-embedding-model/. Accessed: 2023-06-03.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Bevan Koopman and Guido Zuccon. 2016. A test collection for matching patients to clinical trials. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 669–672.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Kevin Lin, Kyle Lo, Joseph E Gonzalez, and Dan Klein. 2023. Decomposing complex queries for tip-of-the-tongue retrieval. *arXiv preprint arXiv:2305.15053*.

Jiduan Liu, Jiahao Liu, Qifan Wang, Jingang Wang, Wei Wu, Yunsen Xian, Dongyan Zhao, Kai Chen, and Rui Yan. 2023. RankCSE: Unsupervised sentence representations learning via learning to rank. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13785–13802, Toronto, Canada. Association for Computational Linguistics.

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*.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2023a. Fine-tuning llama for multi-stage text retrieval. *arXiv preprint arXiv:2310.08319*.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023b. Zero-shot listwise document reranking with a large language model. *arXiv preprint arXiv:2305.02156*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718.

Kaihang Pan, Juncheng Li, Hongye Song, Hao Fei, Wei Ji, Shuo Zhang, Jun Lin, Xiaozhong Liu, and Siliang Tang. 2023. Controlretriever: Harnessing the power of instructions for controllable retrieval. *arXiv preprint arXiv:2308.10025*.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, et al. 2023. Large language models are effective text rankers with pairwise ranking prompting. *arXiv preprint arXiv:2306.17563*.

Devendra Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3781–3797, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2022. Scirepeval: A multi-format benchmark for scientific document representations. *arXiv preprint arXiv:2211.13308*.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121, Toronto, Canada. Association for Computational Linguistics.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agent. *arXiv preprint arXiv:2304.09542*.

Katherine Thai, Yapei Chang, Kalpesh Krishna, and Mohit Iyyer. 2022. RELiC: Retrieving evidence for literary claims. In *Proceedings of the 60th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7500–7518, Dublin, Ireland. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to bm25 and language models examined. In *Proceedings of the 2014 Australasian Document Computing Symposium*, pages 58–65.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, Melbourne, Australia. Association for Computational Linguistics.

Jianyou Wang, Kaicheng Wang, Xiaoyue Wang, Prudhviraj Naidu, Leon Bergen, and Ramamohan Paturi. 2023. Doris-mae: Scientific document retrieval using multi-level aspect-based queries. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Berdersky. 2023. Beyond yes and no: Improving zero-shot llm rankers via scoring fine-grained relevance labels. *arXiv preprint arXiv:2310.14122*.

A Experiment Details

Unless otherwise stated, we use the GPT4-06-13 checkpoint as the LLM and access it via OpenAI's API. Each LLM-based system takes less than 12 hours to run, and costs approximately $500-$1000. All embedding models have 350M parameters or less. Encoder-decoder or decoder models have 7B parameters or less, and experiments on these models can be run on one node of an 8 NVIDIA H100 GPU (80G) within one hour.

A.1 Baselines

For pretrained embedding models, we choose several recent state-of-the-art models as well as several that have been extensively benchmarked on other tasks: E5-v2-Large (Wang et al., 2022), SimCSE-Large (Gao et al., 2021), SPECTER-v2 (Singh et al., 2022), ROBERTA-Large (Liu et al., 2019), SPLADE-v2 (Formal et al., 2021), and SPLADE++ (Formal et al., 2022). There are several encoder-decoder models specifically trained for information retrieval: monoT5 (Nogueira et al., 2020), which is supervised on MS MARCO, and TART (Asai et al., 2023), which is instruction-tuned and can incorporate task descriptions. For DORIS-MAE, ArguAna and WTB, we also report results for E5-L-v2 fine-tuned using synthetic data generated by Promptagator (Dai et al., 2022). Additionally, we evaluate a decoder-only model Rank-LLaMA-7B (Ma et al., 2023a) trained on MS MARCO.

A.2 Full Experiment Results

See Table 3 for full experiment results with more embedding models.

B BIRCO Descriptions

We provide a more detailed description for each dataset in BIRCO. Also see Figure 1.

DORIS-MAE

60 queries that are complex research questions from computer scientists. The query communicates specific requirements from research papers. Each query has a candidate pool sized approximately 110.

ArguAna

100 queries, each with a candidate pool of around 50 passages. Queries and passages are both complex one-paragraph arguments about current affairs. The objective is to find matching counter-arguments.

Clinical-Trial

Model	DORIS-MAE		ArguAna		WTB		Clinical-Trial		RELIC	
	nDCG@10	R@5	nDCG@10	R@5	nDCG@10	R@5	nDCG@10	R@5	nDCG@10	R@5
Embedding Models										
RoBERTa-L	66.8 ± 1.3	12.0 ± 1.6	31.5 ± 3.6	40.2 ± 5.1	14.6 ± 2.7	16.0 ± 3.7	25.9 ± 2.2	8.9 ± 1.2	8.4 ± 2.2	8.9 ± 2.7
SPLADE++	66.8 ± 1.3	8.2 ± 1.2	34.0 ± 3.4	47.9 ± 5.0	9.5 ± 2.2	8.9 ± 2.9	27.5 ± 2.2	9.5 ± 1.2	11.4 ± 2.2	13.9 ± 3.3
SPLADE-v2	67.9 ± 1.4	10.6 ± 2.1	37.8 ± 3.7	40.7 ± 5.0	16.2 ± 3.0	20.2 ± 4.1	22.3 ± 2.3	7.2 ± 1.1	10.7 ± 2.1	11.0 ± 3.0
SPECTER-v2	71.4 ± 1.2	13.5 ± 2.3	37.7 ± 3.3	49.1 ± 5.1	9.8 ± 2.0	13.0 ± 3.4	30.4 ± 2.1	11.2 ± 1.2	7.7 ± 1.6	12.0 ± 3.1
E5-L-v2	72.0 ± 1.2	14.6 ± 1.9	43.5 ± 3.2	62.0 ± 4.8	36.6 ± 4.0	39.9 ± 4.8	29.4 ± 2.7	10.5 ± 1.7	11.1 ± 2.6	14.9 ± 3.4
SIMCSE	70.8 ± 1.3	14.4 ± 1.8	39.8 ± 3.5	51.1 ± 5.1	31.7 ± 4.0	37.0 ± 5.0	27.7 ± 2.7	10.5 ± 1.6	12.9 ± 2.5	15.6 ± 3.7
Promptagator	76.0 ± 1.4	15.6 ± 1.8	63.0 ± 3.3	**77.8** ± 4.2	59.1 ± 4.2	62.9 ± 4.7	NA	NA	NA	NA
Encoder-Decoder Models										
TART	64.9 ± 1.5	10.2 ± 1.7	47.6 ± 3.1	63.7 ± 4.8	42.7 ± 3.7	46.1 ± 4.9	32.8 ± 2.5	10.8 ± 1.3	9.4 ± 1.9	7.9 ± 2.7
TART+O	49.6 ± 1.8	4.9 ± 1.0	23.1 ± 2.8	28.0 ± 4.4	17.9 ± 2.4	20.7 ± 4.1	23.6 ± 2.2	7.7 ± 1.0	14.3 ± 2.3	18.3 ± 3.8
MonoT5	66.9 ± 1.5	12.4 ± 1.8	45.9 ± 3.2	52.0 ± 4.9	50.8 ± 4.2	59.0 ± 4.9	33.2 ± 2.5	**14.2** ± 2.3	13.9 ± 2.7	11.9 ± 3.2
Decoder-Only Model										
Rank-LLaMA	75.1 ± 1.3	**18.4** ± 2.0	53.2 ± 2.9	71.9 ± 4.4	63.7 ± 3.9	69.8 ± 4.6	29.7 ± 2.2	9.8 ± 1.1	15.4 ± 2.4	19.0 ± 4.0
Comparison-based LLM IR Systems										
Rank_{GPT}	76.2 ± 1.3	**17.5** ± 2.3	26.7 ± 2.2	17.0 ± 3.8	**79.5** ± 3.5	**85.9** ± 3.4	39.9 ± 2.6	**14.8** ± 1.4	40.1 ± 4.0	48.0 ± 5.1
$\text{Rank+O}_{\text{GPT}}$	77.4 ± 1.3	**17.6** ± 2.1	54.4 ± 3.2	70.9 ± 4.5	**82.1** ± 3.3	**88.0** ± 3.2	38.6 ± 2.8	14.6 ± 1.5	**62.3** ± 3.9	**70.1** ± 4.5
Comparison-based LLM IR Systems										
$\text{Score+O}_{\text{GPT}}$	**79.9** ± 1.2	**19.3** ± 2.0	51.6 ± 2.7	70.0 ± 4.6	**83.3** ± 3.1	**90.9** ± 2.8	**43.4** ± 2.4	**17.2** ± 1.6	**54.1** ± 3.3	**70.1** ± 4.6
$\text{Reason+O}_{\text{GPT}}$	74.9 ± 1.3	**18.1** ± 2.3	59.9 ± 2.9	**76.8** ± 4.2	74.9 ± 3.6	82.8 ± 3.7	**45.7** ± 3.1	**17.0** ± 1.7	39.9 ± 3.3	51.0 ± 5.0
$\text{Subtask+O}_{\text{GPT}}$	**78.5** ± 1.2	**18.5** ± 1.9	**69.5** ± 3.0	**85.9** ± 3.5	**79.6** ± 3.5	**85.9** ± 3.3	**43.5** ± 2.9	**16.5** ± 1.6	53.2 ± 3.5	**62.9** ± 4.8

Table 3: Experiment results for all models. nDCG@10 and Recall@5 for the benchmark datasets. Bold indicates $p > 0.05$ compared to the highest numerical value indicated in red. The notation +O indicates task objective awareness.

100 queries that are paragraph-length patient case-reports. Each query has a candidate pool comprising 30-110 passages that are paragraph-length descriptions of clinical trials. The objective is to find the most suitable clinical trial for a patient.

WhatsThatBook

100 queries, with each query describing a book in an ambiguous manner. Each query has a pool of 50 passages, which are book descriptions.

RELIC

100 queries which are excerpts from scholars analyzing classic English-language literature. Passages are sentences from a novel that have been extracted from the queries. The objective is to match a literary analysis with its missing quotations.

C Task Objectives for BIRCO

We report our prompts, which were optimized on a small-scale development set for each dataset. The dev set for DORIS-MAE has 40 queries. The dev set for Clinical-Trial has the rest 9 queries. All the other datasets have 50 queries in their dev sets.

C.1 Task Objective for ArguAna

"This information retrieval (IR) task has a debate format where a topic is given, and two directly opposing sides of arguments about this topic are formed. A query is an argument that takes one side of this topic, focuses on a particular point about this topic, and takes a stance (i.e., opinion, position, view, perspective) about this particular point. A passage is an argument that takes the opposing side of the same topic, focuses on the same particular point about the same topic, and takes a directly opposing stance that directly (i.e., no implying or inferring) refutes and attacks the query's stance regarding this particular point. Both query and passage might have citations in them but these citations should not be considered in the scope of this task. The overall goal of this specific information retrieval IR task is to identify the central topic of the debate, to articulate the query's stance, and to find the passage that takes the opposing stance."

C.2 Task Objective for DORIS-MAE

"The query consists of users' needs, leading to several research questions that span a paragraph. Each candidate passage is an abstract from a scientific paper. The objective of this information retrieval task is to identify the abstract that most effectively meets the user's needs in the query."

C.3 Task Objective for WTB

"The query has this format: a user is trying to remember the name of a specific book. The user only remembers some details about the book, such as places, events, and some characters' names. Some of the details are described using informal language. The passage is a book description or summary of a specific book. The passage typically describes the overall storyline of the book and contains some details about the book. The objective of this information retrieval IR task is for you to find the passage that has details or components that holistically best match, explicitly or implicitly, the details or components raised in the query. In other words, you need to find the book description (i.e., the passage) that is most likely the book the user is looking for in the query."

C.4 Task Objective for RELIC

"The query is a piece of literary analysis written by a scholar. In the query (i.e., the excerpt from a literary analysis), one or more quotations from a classic English novel is used as evidence to support the claims made by the literary analysis. Quotations are identified by quotation marks. Now, one quotation is being intentionally masked from the literary analysis (i.e., the query), and replaced by the symbol [masked sentence(s)]. An important claim is made in the preceding context and another important point is made in the subsequent context surrounding the [masked sentence(s)]. The objective of this information retrieval task is to find the most suitable passage that can be used to **directly support** at least one claim made in the query (i.e., the claim that is made in the preceding or the claim subsequent context surrounding the [masked sentence(s)]) and is very natural to be plugged into the [masked sentence(s)] part of the query. Obviously the most suitable passage should **NOT REPEAT** or be contained in any part of the query. It does not make sense to repeat the same or very similar things twice in literary analysis."

C.5 Task Objective for Clinical-Trial

"The motivation of the Information Retrieval task is that clinical trials are experiments conducted in the development of new medical treatments, drugs or devices, and recruiting candidates for a trial is often a time-consuming and resource-intensive effort. A query is a patient case report (either in the form of electronic patient records or ad-hoc queries). A passage is a clinical trial. This Information Retrieval task is to improve patient recruitment for clinical trials. The overall goal of this specific information retrieval IR task is to match eligible patients (the query) to clinical trials (the passage) for recruitment."

D Prompt Structure for Score and Reason

Please refer to Figure 4, 5 for the prompts for Reason+O_{GPT} and Subtask+O_{GPT}.

E The Effects of Candidate Pool Construction

Please refer to Table 4 for the statistics about the test set and the whole dataset.

F Example (Query, Passage) Pairs from the Dataset

Please refer to Figure 6, 7, 8, 9, 10 for examples of query and passage pairs from the datasets.

In an information retrieval task setting, you have a query and a candidate pool of passages. **{task description}**

Now, you are presented with a query and a single passage from its candidate pool. Given the above IR task description, determine the appropriate criteria to make your decision about whether this passage is what the IR task is looking for. After you have written your decision criteria, you need to follow this decision criteria by thinking step by step, and then output a single score on a scale of 0-10. A lower score would mean the passage is less likely to be the passage the IR task wanted, whereas a higher score would mean the passage is more likely to be the passage the IR task wanted. In addition, after the symbol "QUOTES:", list all *direct quotes*, in a python list, verbatim from the passage that definitively suggests or supports why this passage is what the IR task is looking for. A high score would potentially indicate a large amount of supportive quotes from the passage being listed whereas a low score would potentially indicate no or very few quotes from the passage being listed. I need to parse your answer in a certain format, so in the end, only output the score in this format. SCORE:

Query: **{query}**

Passage: **{passage}**

Figure 4: Prompt for Reason+O_{GPT}

Here is the passage: **{passage}**

Here is the requirement extracted from the query: **{req}**

Here is the query: **{query}**

You are a helpful assistant. A user wants to find a passage that satisfies a particular requirement from a query. You are shown the query so that you understand this particular requirement's context, but your objective is to focus solely on the requirement and the passage. Based on the passage, you will determine whether the requirement is met.

Figure 5: Prompt for Subtask+O_{GPT}

Dataset	Test Set		Whole Dataset	
	E5-v2 nDCG@10	E5-v2 R@5	E5-v2 nDCG@10	E5-v2 R@5
ArguAna	43.5 $\pm$ 3.2	62.0 $\pm$ 4.8	48.0 $\pm$ 1.0	59.7 $\pm$ 1.4
WTB	36.6 $\pm$ 4.0	39.9 $\pm$ 4.8	21.0 $\pm$ 2.8	24.7 $\pm$ 3.5
Clinical-Trial	29.4 $\pm$ 2.7	10.5 $\pm$ 1.7	32.4 $\pm$ 2.6	13.1 $\pm$ 1.9
RELIC	11.1 $\pm$ 2.6	14.9 $\pm$ 3.4	8.2 $\pm$ 0.4	10.1 $\pm$ 0.5

Table 4: Model performance on BIRCO's test sets and the full original datasets

Query:

I am working on autonomous driving research and I need to minimize the risk of autonomous vehicles to the public, especially pedestrians. Therefore, I am trying to build a model that will inform my autonomous vehicles to avoid crowded streets where lots of people are walking. This model has to be deployed in real-time. My current idea is to use a computer vision model that can extract and count the number of pedestrians from surveillance footage videos. It needs to be a fast pedestrian detection model. How should I go about doing this? Should my model be able to process videos or simply just images ? If I use videos, the computational load would be higher but I could get a more accurate estimate of the walking speed of these pedestrians during a time interval.

Passage with Highest Relevance Score:

Detecting and predicting the behavior of pedestrians is extremely crucial for self-driving vehicles to plan and interact with them safely. Although there have been several research works in this area, it is important to have fast and memory efficient models such that it can operate in embedded hardware in these autonomous machines. In this work, we propose a novel architecture using spatial-temporal multi-tasking to do camera based pedestrian detection and intention prediction. Our approach significantly reduces the latency by being able to detect and predict all pedestrians' intention in a single shot manner while also being able to attain better accuracy by sharing features with relevant object level information and interactions.

Figure 6: Example from DORIS-MAE

Query:

There is no good solution for the crisis Greece finds itself in, only less bad ones. Austerity measures imposed on Greece may currently be causing suffering, but austerity is the least bad option available for the Greek people: default would be considerably worse. Here is what would most likely happen: The Greek banking sector would collapse [1]. A large portion of the Greek debt is owed to Greek banks and companies, many of which would quickly go bankrupt when the Government defaults. This is also because Greek banks are almost totally reliant on the ECB for liquidity. [2] People would consequently lose their savings, and credit would be close to impossible to find. The Government would quickly devalue the Drachma by at least 50%. This will lead to imported goods being more expensive and consequently to a huge rise in inflation with the living costs increasing tremendously.[3] These two events would lead to a severe shortage of credit, making it almost impossible for struggling companies to survive. Unemployment would soar as a result. It will become increasingly difficult to secure supplies of oil, medicine, foodstuffs and other goods. Naturally, those hit worst would be the poor. The Government, in this respect, would be failing on an enormous scale in providing many citizens with the basic needs. [4] [1] Brzeski, Carsten: “Viewpoints: What if Greece exits euro?”, BBC News, 13 July 2012, [2] Ruparel, Raoul and Persson, Mats: “Better off Out? The short-term options for Greece inside and outside of the euro”, June 2012, Open Europe, 2012 [3] ibid [4] Arghyrou, Michael: “Viewpoints: What if Greece exits euro?”, BBC News, 13 July 2012,

Passage with Highest Relevance Score:

It is not necessarily true that the whole banking sector in Greece would collapse. Given that the default would be orderly and take place within the context of the European Union, the ECB and European Commission would still provide substantial liquidity aid for Greek banks. Moreover it is not true that a devaluation of domestic currency necessarily leads to high inflation – this was not the case, for example, when Britain exited the European Exchange-rate Mechanism in 1992 and pursued a devaluation policy of the British Pound. [1] Lastly, evidence of recent governments that have defaulted suggests that even though some of the harms the opposition refer to may actualise, recovery generally follows fairly quickly, as was the case with Argentina, South Korea and Indonesia. [2] [1] Ruparel, Raoul and Persson, Mats: “Better off Out? The short-term options for Greece inside and outside of the euro”, June 2012, Open Europe, 2012 [2] Becker, Garry: “Should Greece Exit the Euro Zone?”, The Becker-Posner Blog, 20.5.2012,

Figure 7: Example from ArguAna

Query:

This must be a golden-age SF short story. I read it as a kid, translated into Russian, probably later 80s or early 90s but the story must be older than that. I remember that some kind of personal confrontation happens on a different planet. The distinctive landscape feature are these big dust-filled sinkholes, in which a person could fall in and get lost forever.For some reason, the hero, and, possibly others, are using plastic flesh or some kind of artificial flesh - I don't remember if this is for impersonating others or for some other reason. In the end the hero wins. He uses the sinkhole to hide in. He is badly burned. The ending is something along the lines, "He gathered all pain from all over his body mentally into a single point, then lifted that point of the body. When they found him, half of his face was horribly burned. The other half was smiling blissfully."

Passage with Highest Relevance Score:

Otto McGavin is peacefully idealistic by nature, an Anglo-Buddhist, who seeks employment with the Confederacion because he believes in its mission to protect human & nonhuman rights. The only problem is that the Confederacion needs him as one of its twelve Prime Operators for its secret service, the TBII. The TBII wants him as a spy, thief & assassin. It's not, of course, Otto McGavin is peacefully idealistic by nature, an Anglo-Buddhist, who seeks employment with the Confederacion because he believes in its mission to protect human & nonhuman rights. The only problem is that the Confederacion needs him as one of its twelve Prime Operators for its secret service, the TBII. The TBII wants him as a spy, thief & assassin. It's not, of course, a problem for the Confederacion, which simply uses immersion therapy & hypnotic personality overlay for Otto's training, then sends him out in deep cover, encased in plastiflesh, on a variety of dangerous missions on a number of bizarre worlds. But for him, it's a different matter: what he has to witness & what he's forced to do take a terrible toll. Always he returns to his original self--his conscience stabbed by the memory of all those he'd killed in the service of interstellar harmony. ...more

Figure 8: Example from WhatsThatBook

Query:

A 44-year-old man was recently in an automobile accident where he sustained a skull fracture. In the emergency room, he noted clear fluid dripping from his nose. The following day he started complaining of severe headache and fever. Nuchal rigidity was found on physical examination.

Passage with Highest Relevance Score:

Adding vancomycin to the antibiotic regimen is recommended for the treatment of pneumococcal meningitis in adults. Use of dexamethasone as adjunct therapy has proved to reduce mortality and neurologic sequelae in adult patients with pneumococcal meningitis. However, use of dexamethasone may impair penetration of vancomycin in cerebrospinal fluid. In a purely observational manner, we thought to measure blood and CSF concentrations of vancomycin in adult patients with pneumococcal meningitis, treated with vancomycin, third-generation cephalosporin and dexamethasone. Because of a considerable increase in streptococcus pneumoniae meningitis with penicillin nonsusceptible strains, it is now largely recommended to add vancomycin to the third-generation cephalosporin antibiotic regimen. It has also been recently shown that use of dexamethasone reduces mortality and unfavorable outcome in adults with pneumococcal meningitis. However, concern has arisen, that dexamethasone may impair penetration of vancomycin in cerebrospinal fluid. We therefore thought to measure in a purely observational study, blood and CSF vancomycin concentrations in adult patients with pneumococcal meningitis hospitalized in medical intensive care unit that received third-generation cephalosporin, vancomycin and dexamethasone. The aim of the study was to observe whether or not sufficient concentrations of vancomycin could be measured in the CSF despite the concomitant use of dexamethasone. Patients were cared for in a perfectly routine manner. There was no randomization. All patients received routine, recommended care (IDSA guidelines). There was no invasive procedure. Dexamethasone was administered according to the de Gans study (NEJM 2002).

Figure 9: Example from Clinical-Trial

Query:

It is "a city within a city, a kind of encapsulated citadel of human society. "
" Like society, too, "that noun of multitude or signifying many, called Todgers's", is a collective whole different from and greater than the sum of its parts, more complex by far than anything Dickens was able to imagine in Nickleby. Its structure mirrors the illogical and mystified structure of society, full of disguised connections and obscure relationships. It stands in a "labryinth," and its "grand mystery" is its cellar **[masked sentence(s)]** Here the social mysteries, as is not usually the ease in the later Dickens, are reassuring rather than ominous, and moreover the social labyrinth opens itself to the practiced mastery of the initiate. "
Congcsted, shabby, haphazard, impenetrable, and withal utterly humanized," Todgers's is, says Marcus, "the visible and palpable presence of a complex civilization and its history, eccentric, elaborate, thick, various, outlandish, absurd. As for the people who live at Todgers's: Dickens's point is not simply that they are wholly at home there, as is emphatically so.

Passage with Highest Relevance Score:

But the grand mystery of Todgers's was the cellarage, approachable only by a little back door and a rusty grating; which cellarage within the memory of man had had no connection with the house, but had always been the freehold property of somebody else, and was reported to be full of wealth; though in what shape-whether in silver, brass, or gold, or butts of wine, or casks of gun-powder-was matter of profound uncertainty and supreme indifference to Todgers's and all its inmates.

Figure 10: Example from RELIC