SynDL: A Large-Scale Synthetic Test Collection for Passage Retrieval

Hossein A. Rahmani
University College London
London, UK
hossein.rahmani.22@ucl.ac.uk

Xi Wang
University of Sheffield
Sheffield, UK
xi.wang@sheffield.ac.uk

Emine Yilmaz
University College London
Alan Turing Institute & Amazon
London, UK
emine.yilmaz@ucl.ac.uk

Nick Craswell
Microsoft
Seattle, US
nickcr@microsoft.com

Bhaskar Mitra
Microsoft
Montréal, Canada
bmitra@microsoft.com

Paul Thomas
Microsoft
Adelaide, Australia
pathom@microsoft.com

Abstract

Large-scale test collections play a crucial role in Information Retrieval (IR) research. However, according to the Cranfield paradigm and the research into publicly available datasets, the existing information retrieval research studies are commonly developed on small-scale datasets that rely on human assessors for relevance judgments — a time-intensive and expensive process. Recent studies have shown the strong capability of Large Language Models (LLMs) in producing reliable relevance judgments with human accuracy but at a greatly reduced cost. In this paper, to address the missing large-scale ad-hoc document retrieval dataset, we extend the TREC Deep Learning Track (DL) test collection via additional language model synthetic labels to enable researchers to test and evaluate their search systems at a large scale. Specifically, such a test collection includes more than 1,900 test queries from the previous years of tracks. We compare system evaluation with past human labels from past years and find that our synthetically created large-scale test collection can lead to highly correlated system rankings.

CCS Concepts

• **Information systems** → **Information retrieval**.

Keywords

Synthetic Data Generation, Large Language Model, Test Collection

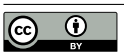
ACM Reference Format:
Hossein A. Rahmani, Xi Wang, Emine Yilmaz, Nick Craswell, Bhaskar Mitra, and Paul Thomas. 2025. SynDL: A Large-Scale Synthetic Test Collection for Passage Retrieval. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3701716.3715311

1 Introduction

Text retrieval approaches identify complete documents relevant to a query. Relevance can be computed based on the similarity of the query and document as determined by comparing the query to words and passages in documents [8]. Alternatively, due to the likelihood of including non-relevant or redundant information in a document and the efficiency of locating relevant information in documents, passage retrieval has become a common research task in information retrieval with the development of many passage retrieval models [17]. Meanwhile, the introduction of effective passage retrieval solutions strongly ties to a well-rounded evaluation. Many existing evaluation operations are commonly instructed by the Cranfield paradigm [4] using a test collection to determine the performance of an information retrieval system. A basic test collection needs to comprise a large set of documents or passages, a set of information needs in plain text, and the corresponding relevance judgments for every document or passage when referring to each information need. Many known and commonly investigated test collections include MS MARCO [11], the collections released by years of organisation of evaluation campaigns like TREC (e.g., TREC Deep Learning Tracks [5, 6]), and workshops or conferences like CLEF and NTCIR.

However, even though there are many available test collections [10], it has been a common concern in the information retrieval community about the shortage of a large-scale test collection for modelling the complex relationships between queries and documents and developing advanced passage and document ranking approaches [7]. Indeed, using MS MARCO as a typical test collection example, it has over 1M of questions that can act as queries. However, for each query, only an average of 10 passages may contain the answer to the query, leaving about 8.8M passages as non-relevant [11]. Similarly, for the test collection of TREC Deep Learning (2023) [6], even though it has richer labels about the query to passages in different relevance levels (related, highly relevant, perfectly relevant), we still observe a low number of queries (i.e, 82) for evaluation, which has been the highest since the tracks from 2019. Hence, it is difficult for a model to capture the complex relationships when modelling the relevance of a query to a large passage corpus, especially in the initial ranking stage.

On the other hand, over recent years, the rapid advancement of machine learning techniques, especially with the introduction of large language models capable of promising natural language comprehension and generation ability, has greatly updated the research development strategies in the information retrieval community.

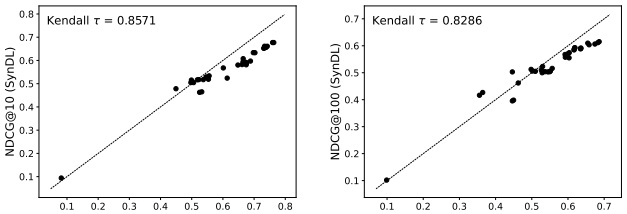


WWW Companion '25, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1331-6/25/04
https://doi.org/10.1145/3701716.3715311

Table 1: SynDL dataset statistics

Data	DL-19	DL-20	DL-21	DL-22	DL-23	SynDL
TREC (Judged) Queries	43	54	53	76	82	**1,988**
TREC (Initial) Queries	200	200	477	500	700	**1,988**
TREC Qrels	9,260	11,386	10,828	386,416	22,327	**637,063**
TREC Qrels/Query	215.3	210.9	204.3	5,084.4	272.2	**320.45**
TREC Docs	8.8M	8.8M	138M	138M	138M	**146.8M**
Irrelevant (0)	5,158	7,780	4,338	286,459	13,866	369,567
Related (1)	1,601	1,940	3,063	52,218	4,372	126,406
Highly relevant (2)	1,804	1,020	2,341	46,080	2,259	86,162
Perfectly relevant (3)	697	646	1,086	1,659	1,830	54,928



(a) NDCG@10 **(b) NDCG@100**

Figure 1: System Ranking correlation test between two test collections (DL-19 and SynDL).

Some examples are the growing research outputs on dense passage retrieval techniques [8], instructed language model for user-intent aware retrieval [2], and Query2Doc [18] that generates pseudo documents based on a query by prompting language models. In particular, following the effectiveness of natural language generation via large language models, we see the potential of language models in making judgments about the relevance between queries and documents [13–15]. Indeed, as discussed in [12], a high correlation between using human and LLM judgments when assessing system rankings has been observed, which encourages the introduction of a large-scale test collection to domains in need. Hence, in this paper, we aim to contribute to the development of a high-quality large-scale passage retrieval test collection with the use of large language models. It is worth noting that large language models could act as a ranking model directly. However, it is essential to adopt large-scale language models to ensure satisfactory performance [7] and its use can be time and economically costly to the retrieval process. In this study, we ground our development by extending the test collections from the five years of the TREC Deep Learning (DL) tracks using large language models to distil the relevance assessment. We refer to the developed passage retrieval test collection with the name of SynDL, its release and many associated baseline approaches from years of TREC DL submissions aim to support in addressing the following research challenges in the community:

- **Deep Relevance Assessment**: The existing passage retrieval test collection often provides few relevance labels on documents for each query, which results in shallow evaluation.
- **Diverse Evaluation**: Many passage test collections, like the ones used in TREC DL tracks, use a small number of queries and limit the evaluation to a small set of test query samples.
- **Rich Baselines**: The inclusion of a small list of baselines often ignores the rich insightful comparisons while introducing novel techniques.
- **Synthetic Query Analysis**: Existing test collections do not enable extensive analysis into the case of comparing synthetic queries and human-provided queries with deep query relevance labels.

In this study, with the notice of the above challenges, we first provide a comprehensive and detailed description of the SynDL test collection. In addition, we augment the introduction of this test collection with extensive evaluations on the alignment of using LLM judgments to human judgments, the comparison of the difference of the resulting system ranking orders and the potential bias effect that might introduced by the use of LLM judgments. With the in-depth evaluation and analysis, we show the high quality of our test collection in providing aligned passage retrieval system rankings to human assessors with "deep and wide" relevance labels.

2 SynDL Test Collection Development

With a focus on the task of passage ranking, we aim to extend the popular test collections of TREC Deep Learning tracks and develop a large-scale test collection, named SynDL, by leveraging LLM judgments to mitigate the discussed research challenges caused by the shortage of a test collection with diversified queries and deep document relevance labels. To illustrate the test collection development process, we first describe the base test collections sourced from the TREC Deep Learning tracks, then followed by the test collection extension strategy with the use of LLM judgments.

The TREC Deep Learning (DL) Track is an initiative organized by the National Institute of Standards and Technology (NIST) to advance the state-of-the-art in information retrieval (IR) and related tasks using deep learning techniques. This track focuses on evaluating the performance of deep learning methods on large-scale datasets and encourages the development of new models and techniques in this domain. The organisation of TREC DL Track was initiated in 2019 [5] and has its final edition in 2023. It has a main focus on two information retrieval tasks, document retrieval and passage retrieval. In particular, each task uses labels provided by human assessors that justify if a passage can answer a given query from the MS MARCO dataset. Note that, in the last run of DL-23, "synthetic queries" are also included in the test collection for non-official evaluation to gain additional insights when compared to the official human evaluations. In Table 1, we present a statistical summary of the TREC Deep Learning test collections over the five years of runs. It is noticeable that, on average, all the test collections rely on a small set of test queries but a reasonable size of relevant documents for performance assessment. However, it is known that the use of small-size test samples could result in inconsistent observation when compared to the use of complex and diversified test samples [16], especially when "wide and shallow" can outweigh "deep and narrow" test collections empirically [3].

In this paper, to improve the resources from existing TREC Deep learning track runs, we propose to leverage the advanced ability of language models to comprehend natural language and assess the relevance between queries and documents [7]. Specifically, the

development of the extended SynDL test collection is organised in three stages: (1) Initial Query Assemble, (2) Assessment Pool Generation and (3) Automatic judgment with LLM. We provide the corresponding descriptions as follows:

(1) **Initial Query Assemble:** For each test collection resource of Deep Learning tracks, it is associated with a set of initial queries, which was meant to be used for human annotators to assess the document relevance for a full set of initial queries. However, only a portion of queries were selected by the human assessor to provide relevant judgments. For example, regarding the test collection sourced DL-19, 200 initial queries were provided but the human assessors left 157 queries unlabelled. Although the following years of runs increased initial queries up to 700, the selected queries for assessment remained on a small scale. To increase the diversity of the queries included in the test collection, we aggregate all initial queries, with a size of 1,988, sourced from the five runs of Deep learning tracks as the initial query inputs for the use of later LLM judgment. It is worth noting that the initial queries of DL-23 also include 500 synthetic queries generated by GPT-4 and T5 models with 250 synthetic queries each [12]. The inclusion of synthetic queries can allow additional bias investigation study of LLM judgments on synthetic queries.

(2) **Assessment Pool Generation:** With the collected initial queries, we follow the common practice in TREC passage retrieval system assessment to prepare the passage pool. In a TREC passage retrieval evaluation, each submission is required to submit a ranked list of passages for each query. Then, the evaluation will be made by selectively considering varied depths of ranked passages with a set of evaluation metrics, such as NDCG@5 and NDCG@10. For the development of diversified relevant passages in our test collection, we embrace the use of the rich submissions among the five runs to collect a depth-10 pool with good coverage of passages in high relevant probabilities. Overall, we use 37, 59, 63, 100 and 35 submissions corresponding to the runs from DL-19 to DL-23. Note that, we will also include these submissions as baselines with their full descriptions in our GitHub repository. After removing the overlapped passages, we obtain a full set of 637, 063 query-passage pairs for relevance assessment. On average, each query is associated with 320.45 passages for relevance annotation.

(3) **Automatic Judgment with LLM:** After having the query-passage pairs ready, we start the annotation of these inputs via large language models. Specifically, with the recently verified advance of GPT-4 [1] in many natural language tasks, we use GPT-4 for this annotation task with a devised prompt and ask the model to provide annotations in a high granularity (i.e., related, highly relevant and perfectly relevant). It is interesting to observe that human annotators are more sensitive to giving perfectly relevant judgments, while GPT-4 give about equal labels of highly relevant and perfectly relevant. Due to the space limit, we include the used prompt in our GitHub[1] repository for reproducible generation.

[1]https://rahmanidashti.github.io/SynDL/

Table 2: Five top-performing submission runs on SynDL

Run	Rank (DL-23)	Run type	NDCG@10	NDCG@100	AP
naverloo-rgpt4	1	prompt	0.9060	0.7841	0.5628
naverloo-frgpt4	2	prompt	0.9007	0.7841	0.5651
naverloo_fs_RR_duo	3	prompt	0.8849	0.7782	0.5590
cip_run_2	4	prompt	0.8671	0.7101	0.4866
cip_run_1	5	prompt	0.8671	0.7101	0.4867

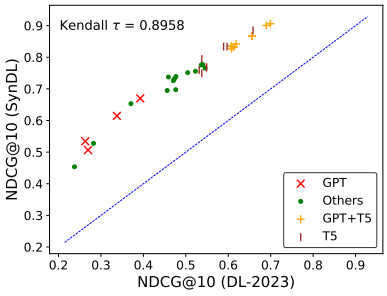


Figure 2: Scatter plots of the effectiveness of DL-23 runs based on SynDL synthetic queries vs. DL-23 test collection to analyse the bias towards systems using the same language model as the one used in synthetic query construction.

After the annotation with LLM judgments, we receive a large-scale test collection with 637,063 query-passage relevance labels and a rich set of queries (1,988). However, as a test collection, it is essential to evaluate the quality of the generated LLM judgments. Hence, we also conduct extensive analysis on the generated test collection for quality evaluation.

3 Resource Evaluation

To effectively evaluate our SynDL test collection, we follow the evaluation setups in [7, 12], which use the correlation test on the system ranking when evaluated using human judgments and LLM judgments. In particular, we compare the performance of systems that were submitted to the five runs of deep learning tracks. Note that, due to the space limit, we only present the correlation test results on DL-19 in this paper and we observe similar results across the comparison on all TREC DL test collections. The rest of the test results will be made available in our GitHub repository for a complete comparison. With the evaluation of the performance of 37 systems that were submitted to DL-19 for official judgment, we compare the ranking difference of these submissions between the use of human assessments and synthetic relevance judgments in our SynDL test collection. Figure 1 shows the evaluated correlation via Kendall rank correlation coefficients when evaluated with NDCG in two depths (@10 and @100). The line of $y = x$ is also included for comparison. According to the value of the correlation test results, we observe a high system ordering agreement using the two test collections with Kendall's τ = 0.8571 and 0.8286 for NDCG@10 and @100, respectively.

In addition to the system ordering agreement evaluation, we also include another comparative study about the agreement on the top-ranked passage ranking systems. The purpose of this evaluation is to observe if the top-ranked systems can also be observed when

evaluated on our SynDL test collection, which can further justify the correlation of the made LLM judgments with the human assessors. In Table 2, we present the evaluation results of top-ranked submissions to the DL-23 on our SynDL test collection. First, regarding the rank difference, we can see that the top 5 systems remain the best-performing passage rankers among the rest of the submissions to DL-23. In particular, the two ranking orders are identical if we consider the NDCG@10 and NDCG@100 measures.

Moreover, with the inclusion of synthetic queries in our SynDL test collection, we further conduct a bias analysis to examine if our test collection would favour systems also using the same or similar language models. As discussed in [9], there is a potential bias towards LLM-generated text when using LLM for evaluation. Hence, to explore this bias effect, we first categorise the submissions to the DL-23 into four categories according to whether they are based on GPT (×), T5(|), GPT + T5 (+) or others (•). Figure **??** shows the system order agreement between the use of our SynDL and DL-23 with highlighted different types of systems. We observe that a high agreement can still be observed between human assessment and language model judgments in this case. GPT-based systems do not get higher ranks when evaluated with GPT-4 generated relevance judgments in our SynDL test collection.

Overall, we experimentally verify that our SynDL test collection is of a high quality, which not only exhibits a high agreement with the human assessors across the comparison to multiple sets of test collections but also shows a robust evaluation outcome when evaluated on the potential bias about using the identical language models.

4 Discussions and Conclusions

In this paper, we summarise the construction of a large-scale test collection (SynDL) with LLM-based relevance judgment for passage retrieval, which is developed based on the test collections from the five runs of TREC Deep Learning tracks. The resulting test collection, SynDL, covers a rich set of queries with deep relevance labels on passages. After conducting a thorough quality evaluation of SynDL, we observe a high agreement between SynDL and every TREC DL test collection on system ordering. In addition, we also highlighted that SynDL with language model relevance judgment does not favour language model approaches according to our conducted experiments.

Recall the observed research challenges for the passage retrieval task. We conclude that our SynDL test collection is promising in providing deep relevance assessment with rich relevance labels and a diverse set of queries. In addition, with the inclusion of passage retrieval systems that were submitted to the TREC DL tracks, we enable the comparison over a rich set of baseline approaches. Moreover, by comparing the research findings on the real and synthetic queries, SynDL also allows extensive research studies to evaluate passage retrieval systems on different types of queries. While our preliminary analysis showed the promising value of our SynDL test collection, we also see the many possibilities of research contributions with its release, such as the transfer learning of using models pre-trained on our test collection, the re-visit of many existing passage retrieval approaches and the development of generalisable passage retrieval techniques on diverse queries.

Acknowledgments

This work is supported by the Engineering and Physical Sciences Research Council [EP/S021566/1], the EPSRC Fellowship titled "Task Based Information Retrieval" [EP/P024289/1], the Turing Fellowship scheme.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] 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*. 3650–3675.

[3] Ben Carterette, Virgil Pavlu, Evangelos Kanoulas, Javed A Aslam, and James Allan. 2008. Evaluation over thousands of queries. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 651–658.

[4] Cyril W Cleverdon. 1960. The aslib cranfield research project on the comparative efficiency of indexing systems. In *Aslib Proceedings*, Vol. 12. 421–431.

[5] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. 2020. Overview of the TREC 2019 deep learning track. In *Trec*.

[6] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Hossein A. Rahmani, Daniel Campos, Jimmy Lin, Ellen M. Voorhees, and Ian Soboroff. 2024. Overview of the TREC 2023 Deep Learning Track. In *Text REtrieval Conference (TREC)*. NIST, TREC. https://www.microsoft.com/en-us/research/publication/overview-of-the-trec-2023-deep-learning-track/

[7] Guglielmo Faggioli, Laura Dietz, Charles LA Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, et al. 2023. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*. 39–50.

[8] Marcin Kaszkiel and Justin Zobel. 1997. Passage Retrieval Revisited. In *Proceedings of SIGIR*.

[9] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[10] Sean MacAvaney, Andrew Yates, Sergey Feldman, Doug Downey, Arman Cohan, and Nazli Goharian. 2021. Simplified data wrangling with ir_datasets. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2429–2436.

[11] 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. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*.

[12] Hossein A Rahmani, Nick Craswell, Emine Yilmaz, Bhaskar Mitra, and Daniel Campos. 2024. Synthetic Test Collections for Retrieval Evaluation. *arXiv preprint arXiv:2405.07767* (2024).

[13] Hossein A Rahmani, Clemencia Siro, Mohammad Aliannejadi, Nick Craswell, Charles LA Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz. 2024. Report on the 1st Workshop on Large Language Model for Evaluation in Information Retrieval (LLM4Eval 2024) at SIGIR 2024. *arXiv preprint arXiv:2408.05388* (2024).

[14] Hossein A. Rahmani, Clemencia Siro, Mohammad Aliannejadi, Nick Craswell, Charles L. A. Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz. 2024. LLM4Eval: Large Language Model for Evaluation in IR. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 3040–3043. doi:10.1145/3626772.3657992

[15] Hossein A Rahmani, Emine Yilmaz, Nick Craswell, Bhaskar Mitra, Paul Thomas, Charles LA Clarke, Mohammad Aliannejadi, Clemencia Siro, and Guglielmo Faggioli. 2024. LLMJudge: LLMs for Relevance Judgments. *arXiv preprint arXiv:2408.08896* (2024).

[16] Mark Sanderson et al. 2010. Test collection based evaluation of information retrieval systems. *Foundations and Trends® in Information Retrieval* 4, 4 (2010), 247–375.

[17] Courtney Wade and James Allan. 2005. Passage retrieval and evaluation. *Tech. Reports of DTIC* (2005).

[18] Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query Expansion with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 9414–9423.