SynSim: Measuring Distributional Realism in Synthetic Query–Passage Datasets for Information Retrieval

Anonymous ACL submission

Abstract

This work addresses the challenge of evaluating synthetic datasets generated by Large Language Models (LLMs) for information retrieval tasks. Generating such data at scale offers many advantages, but human evaluation is costly and time-consuming. We propose an automatic metric that measures the realism of synthetic datasets by comparing the distribution of query–passage similarity scores with those from human-authored datasets. Our results show that the metric captures improvements produced by prompt-engineering techniques, which have already been validated in other contexts, demonstrating its reliability. Finally, the findings are corroborated through both LLM-as-a-judge assessments and human evaluation on a subset of the data, confirming the metric's effectiveness in reflecting meaningful quality differences.

1 Introduction

Recent advances in pre-trained language models, particularly BERT-based models, have enabled a wide range of semantic similarity tasks. These models effectively capture relationships between texts, but they require large, domain-specific datasets and substantial computational resources. Expert annotation produces high-quality data but is often costly and infeasible at scale (Orr and Crawford, 2024). As a scalable alternative, large language models (LLMs) such as ChatGPT, LLaMA, and Mistral can generate synthetic datasets for question-answering and passage retrieval. However, evaluating the quality of synthetic datasets remains a challenge: human evaluation is resource-intensive, automatic metrics often approximate quality, and LLM-as-a-judge methods can introduce bias or be computationally expensive (Wang et al., 2024).

To address these challenges, we propose a novel automatic evaluation framework that represents each dataset as a distribution of semantic similarity scores between queries and their associated passages. By comparing the distributions of synthetic and original datasets, our approach captures dataset-level quality, including the coverage of highly similar and dissimilar query–passage pairs and potential hallucinations (Shumailov et al., 2024). Synthetic datasets are generated using prompt-engineering techniques inspired by InPars (Bonifacio et al., 2022) and further enhanced by providing explanations for poorly formulated questions. This improves synthetic dataset quality, enhances retrieval performance, and reduces the risk of dataset shift (Shuang and Pozi, 2024).

Validation is performed using both an LLM-as-judge framework and human evaluation on a subset of the data. The results show that our distributional similarity metric reliably reflects improvements in synthetic dataset quality and provides a scalable, reproducible evaluation methodology.

Research Questions. Our study addresses two key questions: (i) whether overlap in similarity distributions of query–passage pairs between real and synthetic datasets correlates with actual dataset quality, and (ii) whether prompt-based approaches that account for question formulation errors can improve synthetic question quality, as measured by our similarity-overlap metric.

2 Background

We will use state-of-the-art techniques related to information retrieval. We will apply the data augmentation idea to textual data, using the power of large language models and their capability to produce text.

2.1 Question-Answering

Question answering (QA) systems aim to retrieve relevant information from one or more documents. Typically, documents are divided into chunks, which are ranked based on their semantic simi-

larity to the user's query. Variations include selecting the most similar chunk, the top-k most similar chunks, or even the least similar ones. For example, Bistarelli et al. (Bistarelli and Cuccarini, 2024) combine classical QA with zero-shot learning techniques. One widely used approach to compute semantic similarity is contextual embeddings, which encode language properties into high-dimensional numeric vectors. The similarity between two sentences is reflected by the spatial distance between their vectors. Models based on BERT (Devlin et al., 2019), particularly Siamese BERT (Reimers and Gurevych, 2019), are commonly used, where two BERT models share weights to estimate sentence similarity effectively.

2.2 Text augmentation

Data augmentation techniques are widely used in the world of machine learning, improving performance and providing robustness to results (Rebuffi et al., 2021). It can be used in classification tasks to balance the data set and increase the sample count of the minority class.

There are a number of techniques to increase the data in the case of textual data (Pellicer et al., 2023). They can be divided into two principal categories: feature and data methods. The feature space methods work on the embedding representation of the textual data. An example is to substitute one word for another that has a similar embedding vector. The alternative is to use data space methods. One example is to use knowledge graphs or ontologies that substitute a single word with synonyms or similar words. Otherwise, it can be a reformulation of the sentence or the document using the capabilities of the large language model.

Before the introduction of large language models (LLMs), text augmentation typically involved substituting individual words. One of the first methods to consider entire sentences was back translation (Edunov et al., 2018). This technique involves translating a text into another language and then translating it back to the original language, leveraging the translation model's ability to recognise semantic equivalence.

2.3 Large Language Model for Augmentation

LLMs in the last few years have deeply influenced the field of NLP. Due to their ability obtained thanks large amount of data on pretraining and parameters.

They show great capability in tasks like text classification and sentiment analysis, but also in more complex tasks that require a deep understanding of context and language. These tasks include, for example, text summarisation (Zhang et al., 2024), question answering (Tan et al., 2023), and sentence paraphrasing (Yadav et al., 2024).

They are able to accomplish various tasks with zero-shot learning (ZSL), where the model is asked to solve some tasks without any fine-tuning and using only the prompt to provide information. An alternative to the ZSL is the one-shot or few-shot learning, where they use one or a few samples for the text generation.

Some articles show how the performance of this kind of task has improved greatly in the case of one or a few examples provided as training (Brown et al., 2020). The results demonstrate the great capability of LLMs for NLP-related tasks. The biggest limitation of LLMs is the difficulty in handling large contexts. In the case of a lot of information provided in the prompt, this increases the possibility of hallucinations and requires more computational power (Liu et al., 2024).

For this reason, LLMs are largely used for data augmentation, and in particular for data generation, due to their ability to rephrase and the few samples required to do this.

2.4 Creation of synthetic dataset of text

Machine Learning and Natural Language Processing (NLP) rely on large amounts of labelled data for both training and testing. Creating these data sets is a challenging task that requires significant resources and human effort, and considering important elements such as accessibility, necessity scale, and recognition cost-benefits (Orr and Crawford, 2024). In contexts where such resources are limited, it becomes necessary to use existing datasets. However, if suitable datasets do not exist, synthetic data generation becomes a necessary alternative. This can be done to increase the number of samples in an existing dataset, for example, to address the problem of class imbalance (Fomin et al., 2024), or it can involve the creation of entirely new datasets. The first approach was semi-automatic, where it involved some human verification (Moustafa et al., 2022).

Specifically for NLP, the revolutionary LLM approach, thanks to its ability to generate text, began to be actively used for the task of text annotation (Jahan et al., 2024) in English and also for

other low-resource languages (Frei and Kramer, 2023) or IR-based systems (Kenneweg et al., 2024). The LLM-based approach avoids the need for human verification during the creation of the data set.

LLM-based methods for the creation of a system of passage retrieval usually provide the passage to an LLM and using the prompt is asked to provide a feasible query (Almeida and Matos, 2024). After that, a filter for the quality of the query is applied.

Some works have found that it can be helpful to use LLM in-domain topics for the generation of the question for the QA system (Maufe et al., 2022). One of the biggest limitations of this approach is that the model may focus the question on marginal elements of the passage, losing the real focus. For this reason, Bonifacio et al. (Bonifacio et al., 2022) propose a method that uses a good and bad sample of the question generated for each sample provided in the prompt.

2.5 Similarity Metrics for Distribution Comparison

In this work, we quantify the similarity between distributions of semantic scores using two well-established statistical metrics: the **Overlap (OVL)** and the **Wasserstein Distance**. Each of these measures captures complementary aspects of distributional similarity, providing a robust framework for evaluation. The overlap coefficient measures the shared area between two probability density functions. Its value lies in the range $[0, 1]$, where 1 indicates perfect alignment and 0 indicates no overlap. OVL is particularly intuitive for assessing the extent to which the probability mass of two distributions coincides. Also known as the Earth Mover's Distance (EMD), the Wasserstein distance quantifies the "cost" of transporting probability mass from one distribution to another. Unlike metrics such as the Jensen-Shannon Divergence or OVL, the Wasserstein distance accounts for the geometry of the underlying space, making it sensitive to the magnitude of shifts between probability masses.

3 Data and model

We use established QA datasets as ground truth for evaluating question generation. Specifically, Natural Questions (NQ) (Kwiatkowski et al., 2019) provides open-domain QA with short and long answers from Wikipedia; HotpotQA (Yang et al., 2018) focuses on multi-hop reasoning across multiple passages; and FiQA-2018 (Maia et al., 2018) is a finance-focused QA dataset with sentiment annotations. Only the development and test splits are used, ranging from 1,200 to 14,000 queries, while large training sets and MS MARCO (Nguyen et al., 2016) are excluded for computational reasons. NQ and FiQA are single-hop datasets, whereas HotpotQA requires multi-hop reasoning, which informs prompt design. For query generation, we use generative LLMs of varying sizes: LLaMA 3.1 (8B) and Gemma (12B, 4B), enabling evaluation in low-resource settings. Sentence similarity is measured using Sentence-BERT via the pre-trained `all-MiniLM-L6-v2` model, which encodes semantic information into embeddings; similar sentences produce nearby vectors, while dissimilar ones are farther apart. This model offers a fast and effective solution for semantic search and similarity tasks.

4 Methods

This work introduces a fully automated method for evaluating the similarity between datasets without requiring human annotation or additional model training. Each dataset is represented through similarity scores between queries and their corresponding passages, and synthetic datasets are generated to preserve these properties. Dataset similarity is quantified using two complementary metrics: Overlap (OVL), which measures the shared range of similarity scores, and Wasserstein Distance, which captures differences in both the shape and location of the distributions. Together, these metrics provide a robust, dataset-level assessment of quality. Experiments are conducted on five sampled datasets, with synthetic queries generated using three prompting strategies: a baseline method, InPars, and an improved InPars method. Similarity scores between queries and passages are computed using BERT-based models, producing distributions for both original and synthetic datasets. Comparing these distributions offers a holistic measure of dataset quality, capturing improvements from more advanced prompting strategies and highlighting potential sources of error. Importantly, low-quality or hallucinated queries are deliberately retained, allowing the evaluation to directly reflect the effect of prompt design. The evaluation also includes graphical analyses of similarity distributions and is validated using an LLM-as-judge framework, complemented by human assessment on a subset of data, ensuring that the metrics accurately reflect

improvements in dataset realism and fidelity.

4.1 Question Generation

The first step of our pipeline focuses on generating synthetic questions from a given passage. This approach is practical and scalable: for real-world dataset creation, it suffices to sample a set of documents with informative passages, from which a complete dataset can then be generated automatically. For **question generation**, we employ a series of **prompt-engineering** strategies and evaluate different methods for generating high-quality questions and answers. We apply these methods to both single-hop datasets (NQ and FiQA) and multi-hop datasets (HotpotQA). The **baseline approach** consists of directly prompting the model to generate a question based on one (single-hop) or two (multi-hop) input passages. To guide this process, we provide 10 examples of passage–question pairs for single-hop generation (Appendix A), and 5 examples consisting of two passages and a corresponding question covering both for multi-hop generation. However, providing only positive examples is often insufficient. The model may produce **low-quality questions**—for instance, questions that are overly specific to irrelevant details or that misinterpret the passage context. To address this issue, we introduce a **second approach** in which each prompt contains both **good and bad examples** (Appendix C). This technique, shown to be effective in prior work, helps the model distinguish between well- and poorly-formulated questions. The good examples illustrate *general, well-structured questions*, while the bad examples demonstrate *overly specific or irrelevant ones*. This contrast helps the model better capture the characteristics of a **high-quality question**. Finally, we extend this method by incorporating the **motivation** behind why a question is considered good or bad. Prior studies have shown that including such **explanatory feedback** can improve model performance. By explicitly stating the reasoning that differentiates good from bad examples, the model is guided toward producing **higher-quality and more contextually relevant questions**. In the case of single-hop generation, the motivation focuses on the *clarity and relevance of the question to the passage*, while for multi-hop generation, it emphasizes the *logical connection between the two passages or concepts*.

4.2 Query generation

Once the synthetic dataset is generated, we employ a BERT-based encoder to extract the semantic similarity between each query and its corresponding passage. Specifically, we use **Sentence-BERT**, which represents questions and passages as pairs of dense, high-dimensional vectors. These vectors encode the semantic properties of the text, such that two texts that are close in the vector space are also similar in meaning, even if their surface forms differ.

To measure similarity between vectors, we adopt cosine similarity, which produces a score in the range of -1 and 1 and is usually normalised to obtain a value between 0 and 1.

4.3 Rationale for Metric Selection

In this work, we evaluate synthetic datasets by comparing the distribution of semantic similarity scores with those of original datasets. We focus on four complementary metrics: Overlap (OVL), Wasserstein Distance (WD), Skewness, and Kurtosis. OVL measures the shared area between distributions, indicating how well synthetic queries cover the real semantic similarity range, while WD quantifies distributional differences, capturing subtle shifts in shape or magnitude. Skewness and kurtosis describe asymmetry and tail heaviness of the distributions, providing supplementary insight. Together, these metrics form a robust framework, with OVL and WD as primary indicators of dataset quality and skewness/kurtosis offering additional perspectives on distribution characteristics.

5 Results

We report distributional statistics for FiQA, HotpotQA, and NQ across different LLMs and query types. Our analysis focuses on overlap (OVL), Wasserstein distance, skewness, and kurtosis, which together capture alignment, divergence, asymmetry, and tail behaviour in similarity score distributions.

The goal is to identify patterns between different prompt techniques and our overlap-based evaluation method, demonstrating the ability of LLMs to assess dataset quality without human judgement. We hypothesise that more advanced prompt techniques produce higher overlap with the original distribution compared to naive approaches. Similarly, larger models are expected to generate synthetic datasets of higher quality, achieving stronger

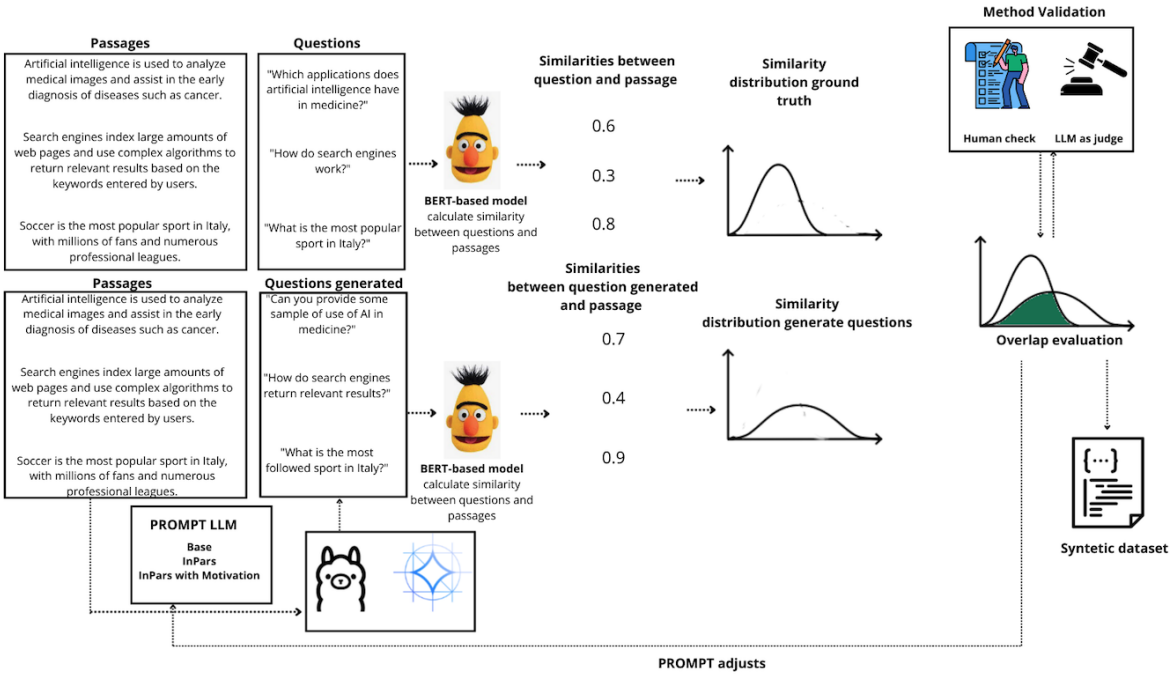


Figure 1: Pipeline description

alignment and higher average overlap. For example, Gemma consistently shows better overlap than Llama, likely due to its larger model size and capacity to capture complex query nuances.

As shown in Table 1, OVL values range from 0.76 to 0.86, indicating moderate to strong overlap between synthetic and original queries. InPars and Motivated prompts improve alignment compared to Base queries. Wasserstein distances remain low (0.045–0.072), indicating limited divergence across query types. Skewness is negative (–0.45 to –1.13), showing left-skewed distributions, while kurtosis varies (2.7–6.1), reflecting occasional extreme similarity scores. Larger models better capture these patterns, highlighting the benefit of prompt-based query generation.

Table 2 shows that HotpotQA exhibits higher overlap than FiQA, with OVL values between 0.90 and 0.94, and lower Wasserstein distances (0.005–0.011), indicating strong alignment between synthetic and original queries. Skewness is mildly negative (–0.55 to –0.63), and kurtosis is near 3, reflecting near-Gaussian, left-leaning distributions. Both Gemma 3:12b and Llama 3.1:8b replicate the similarity structure of the original dataset effectively.

As shown in Table 3, across NQ, InPars queries consistently achieve the highest overlap (0.897–0.944) and lowest Wasserstein distances

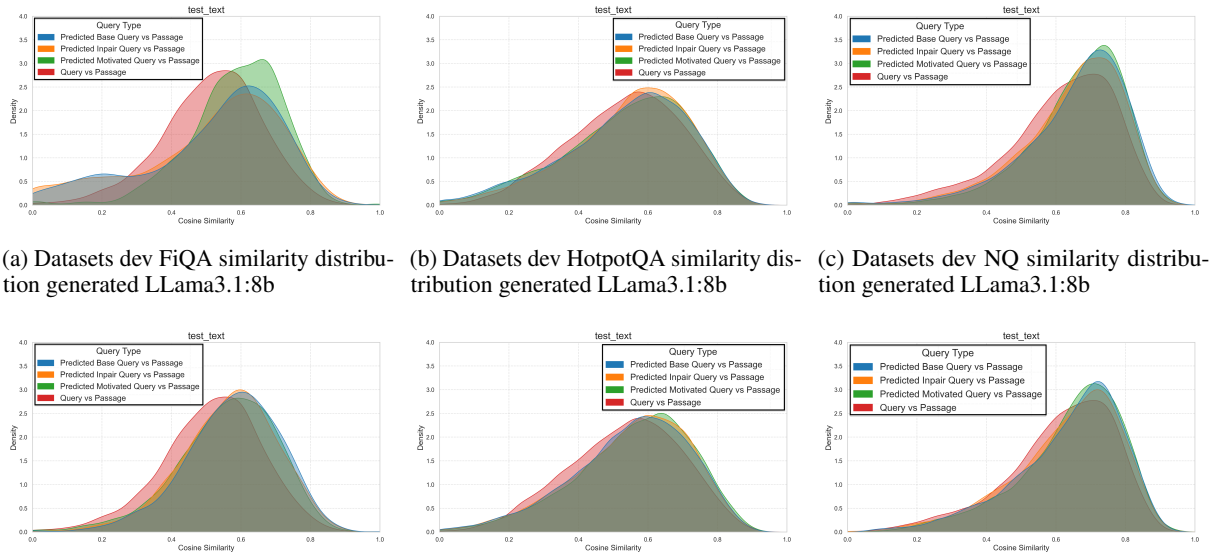
LLM	Split	Query Type	OVL↑	WD↓	Skew	Kurt
gemma3_12b	test	Base	0.8251	0.0612	-0.5309	3.5599
gemma3_12b	test	InPars	0.8535	**0.0513**	-0.4527	3.3610
gemma3_12b	test	Motivated	**0.8605**	0.0454	-0.6435	3.6604
gemma3_12b	dev	Base	0.8292	0.0615	-0.4684	3.5647
gemma3_12b	dev	InPars	0.8476	0.0512	-0.6276	3.9860
gemma3_12b	dev	Motivated	**0.8626**	**0.0482**	-0.4512	3.3770
llama3.1_8b	test	Base	0.7763	**0.0547**	-0.8225	2.9201
llama3.1_8b	test	InPars	0.7798	0.0572	-0.7936	2.8832
llama3.1_8b	test	Motivated	**0.7802**	0.0722	-0.8588	4.7216
llama3.1_8b	dev	Base	0.7683	**0.0562**	-0.8308	3.0526
llama3.1_8b	dev	InPars	0.7604	0.0630	-0.7726	2.7304
llama3.1_8b	dev	Motivated	**0.7906**	0.0693	-1.1320	6.0942

Table 1: FiQA statistics across query generation strategies. Columns report Overlap (OVL), Wasserstein Distance (WD), skewness, and kurtosis.

(0.016–0.034), demonstrating that prompt-based query generation better captures the similarity distribution of the original dataset. Skewness remains negative across models, while kurtosis indicates moderately heavy tails, particularly for larger models. Gemma3:12b generally produces the most aligned synthetic distributions, consistent with its larger size and greater modelling capacity.

6 Error Analysis and Relation Between Original and Generated Queries

An empirical analysis across the dataset compared the three generation strategies, *base*, *InPars*, and



(a) Datasets dev FiQA similarity distribution generated LLama3.1:8b

(b) Datasets dev HotpotQA similarity distribution generated LLama3.1:8b

(c) Datasets dev NQ similarity distribution generated LLama3.1:8b

(d) Datasets dev FiQA similarity distribution generated Gemma3:12b

(e) Datasets dev HotpotQA similarity distribution generated Gemma3:12b

(f) Datasets dev NQ similarity distribution generated Gemma3:12b

Figure 2: The nine sub figure represent the possible distribution of the similarity between the pair query-passage, of the original dataset and the three dataset generated with the prompt approach

LLM	Split	Query Type	OVL↑	WD↓	Skew	Kurt
gemma3_12b	test	Base	**0.9306**	**0.0216**	-0.5674	3.0313
gemma3_12b	test	InPars	0.9219	0.0230	-0.5806	3.0001
gemma3_12b	test	Motivated	0.9094	0.0297	-0.5841	3.0412
gemma3_12b	dev	Base	0.9358	0.0137	-0.5586	2.9197
gemma3_12b	dev	InPars	**0.9399**	**0.0135**	-0.5463	2.9698
gemma3_12b	dev	Motivated	0.9255	0.0198	-0.5758	3.0037
llama3.1_8b	test	Base	0.9237	0.0186	-0.6072	2.9234
llama3.1_8b	test	InPars	**0.9253**	0.0197	-0.6246	3.0572
llama3.1_8b	test	Motivated	0.9244	**0.0178**	-0.5674	2.8334
llama3.1_8b	dev	Base	0.9362	0.0142	-0.5959	2.8898
llama3.1_8b	dev	InPars	0.9265	0.0150	-0.6058	2.9079
llama3.1_8b	dev	Motivated	**0.9395**	**0.0127**	-0.5622	2.8198

Table 2: HotpotQA statistics across query generation strategies. Columns report Overlap (OVL), Wasserstein Distance (WD), skewness, and kurtosis.

LLM	Split	Query Type	OVL↑	WD↓	Skew	Kurt
gemma3_12b	test	Base	0.9193	0.0236	-0.9317	3.6844
gemma3_12b	test	InPars	0.9441	**0.0159**	-0.8665	3.5448
gemma3_12b	test	Motivated	0.9113	0.0288	-0.9216	3.6918
llama3.1_8b	test	Base	0.8652	0.0433	-1.1820	4.9034
llama3.1_8b	test	InPars	**0.8972**	**0.0343**	-1.0623	4.4771
llama3.1_8b	test	Motivated	0.8653	0.0470	-1.1604	5.0941

Table 3: Statistics for the NQ dataset across query generation strategies. Columns report Overlap (OVL), Wasserstein Distance (WD), skewness, and kurtosis.

motivation, focusing on their relation to the original query and their alignment with the corresponding passage. The results show distinct behavioural patterns (Appendix D). The *InPars* model produced outputs that were most lexically and semantically similar to the original queries, often reproducing them with minimal variation. For instance, in "Where did they film the show The Crossing?", all generated forms from *InPars* were identical to the original, demonstrating high surface fidelity. This pattern was recurrent across the dataset and explains why *InPars* achieved the highest similarity and textual support scores. However, this approach offered limited reformulation capacity and little semantic enrichment.

The *base* model displayed a more transformative behaviour, typically reformulating the question by extracting concrete factual cues from the passage. In "theory for functional group test for aldehyde and ketones", *base* generated "what type of precipitate does an aromatic carbonyl compound produce", grounding the query in specific textual evidence about the 2,4-DNPH reaction. This led to a better factual alignment but a narrower conceptual scope. Similarly, for "who took over for Batman when his back was broken", *base* maintained referential precision while simplifying the focus.

Conversely, the *motivation* model produced semantically richer and more explanatory queries, such as "the theory behind the functional group test for aldehyde and ketones using 2,4-Dinitro...". These formulations capture the theoretical rationale behind the question but occasionally introduce de-

Dataset / Query Type	[1]	[2]	[3]	[4]	[5]
Test / Base	4.53	4.12	4.41	2.87	3.85
Test / InPars	4.41	**4.14**	4.35	2.87	3.80
Test / InPars Motivation	**4.67**	4.10	**4.59**	**2.90**	**3.93**
Dev / Base	4.50	**4.13**	4.39	**2.87**	3.85
Dev / InPars	4.46	4.12	4.37	**2.87**	3.82
Dev / InPars Motivation	**4.69**	4.12	**4.62**	**2.88**	**3.95**

Table 4: Predicted metrics for Test and Dev datasets generated by LLama3.1. Column descriptions: 1 Relevance, 2 Clarity, 3 Answerability, 4 Difficulty, 5 Overall score. Query types: Base, InPars, InPars Motivation. Best value in each row is highlighted in bold.

tails not explicitly mentioned in the passage, reflecting interpretive expansion rather than extraction.

Overall, *InPars* maximizes surface similarity, *base* ensures factual grounding, and *motivation* enhances conceptual depth at the cost of precision. The relation between the original and generated queries indicates that preserving semantic intent, not merely lexical form, is key to maintaining both faithfulness and interpretive quality. A hybrid strategy, combining *base*'s textual grounding with *motivation*'s conceptual elaboration, emerges as the most effective balance between accuracy and depth.

7 LLM as judge

To validate our dataset quality metric (OVL and Wasserstein Distance), we use a Large Language Model (LLM) as an automated judge. The LLM rates question–passage pairs across seven dimensions—relevance, clarity, answerability, difficulty, completeness, understanding, and generalisation—on a 1–5 Likert scale, with structured prompts including passages, questions, instructions, and few-shot examples. Outputs are parsed and aggregated to produce dataset-level statistics. Evaluation focuses on the FiQA dataset generated by LLaMA 3.1. This approach provides scalable, consistent, and fine-grained assessment, with LLM scores highly correlated with human judgements (Bubeck et al., 2023; Achiam et al., 2023), confirming the metric's reliability. The LLM is used solely for validation, not for prompt optimization.

8 Graphical evaluation

Figure 2 illustrates that the distributions are strongly influenced by both the dataset and the model type. For the Llama model (Figure 2a), distribution shapes vary across approaches, whereas this variation is less pronounced in Figures 2b and c. In contrast, for the Gemma model, distributions are generally more similar and shifted to the right relative to the original data, indicating greater similarity in generated query-passage pairs. For the HotspotQA dataset (Figures 2b and e), distributions are more stable and closer to the original formulation, likely due to the multi-hop approach, which constrains question generation by requiring involvement of multiple passages. Figures 2c and f show pronounced peaks with fewer high-similarity pairs, suggesting that the original dataset contains inherently more similar query-passage pairs.

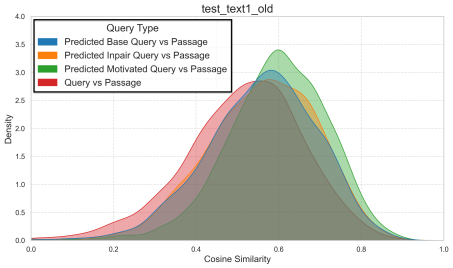
Beyond descriptive insights, these similarity distributions provide a practical evaluation method. By analysing the patterns, one can assess the quality of generated query-passage pairs and guide prompt design—for instance, providing instructions to produce less similar pairs when needed. Leveraging semantic embeddings via the Sentence Model enables this evaluation efficiently, offering a computationally lightweight alternative to the LLM-as-judge paradigm and supporting prompt refinement in a controlled, data-driven manner.

9 Prompt adjustment

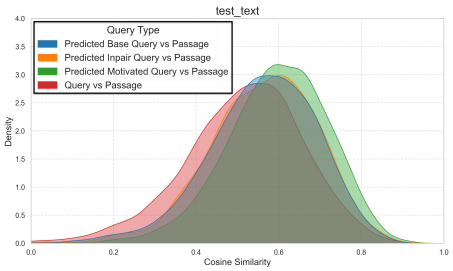
In Figure 3a, the distribution generated by the synthetic dataset is slightly shifted to the right compared to the real distribution, indicating that on average the original pairs are more similar to each other than in the real dataset. This is expected, as LLMs tend to remain closely aligned with the original text when provided with few examples. By refining the prompt to explicitly ask the model to rephrase questions using different wording, the effect can be seen in Figure 3b, where the distribution peak for the InPars Motivated becomes broader and smoother, indicating greater variability, but it does not produce an important effect on the other two approaches. These changes are reflected in the results: for the original prompt, the Base and InPars queries show high similarity to the passage (0.8795 and 0.8758, respectively), while the Motivated queries are lower (0.7617). After prompt refinement (`test_text`), the Base and InPars similarities decrease slightly (0.8707 and 0.8695), while the Motivated queries increase slightly (0.7665).

10 Related work

This work builds on prior efforts in synthetic query generation for information retrieval. InPars (Bonifacio et al., 2022) and InPars Light (Boytsov et al.,



(a) Distribution generated by Gemma3:4b with the prompts base, InPars and InPars Motivated before adjustment



(b) Distribution generated by Gemma3:4b with the prompts base, InPars and InPars Motivated after adjustment

Figure 3: Similarity distribution plots for two datasets generated by LLama3.1:8b.

2024) generate high-quality synthetic queries for passage retrieval, evaluated primarily via BM25. Braga et al. (Braga et al., 2024) extend LLM-based generation to answers using dataset metadata, while LIQUID (Lee et al., 2023) creates multiple-choice question datasets via summarization and entity extraction. Rahmani et al. (Rahmani et al., 2024) demonstrate that LLM-generated queries and relevance judgments can approximate human data. Evaluation methods based on distributional similarity—such as Word Mover's Distance (WMD) (Kusner et al., 2015), BERTScore (Zhang et al., 2020), and embedding-based analyses—show that semantic similarity captures dataset quality more effectively than surface-level overlap. Prior work also explored dataset-level distributional shifts: Kramchaninova et al. (Kramchaninova and Defauw, 2022) studied multilingual QA, and Shumailov et al. (Shumailov et al., 2024) showed that recursive synthetic training induces shifts. In contrast to these approaches, our method emphasizes dataset-level evaluation for synthetic QA, quantifying realism by measuring the overlap between semantic similarity distributions of real and generated query–passage pairs, providing a holistic view of dataset quality beyond individual instance metrics.

11 Future works

The work outlines several directions for future research. It suggests expanding beyond BERT-based semantic analysis to evaluate more aspects of the query-passage relationship and improve result presentation. Future efforts include implementing automated feedback mechanisms for self-assessment, using targeted or random sampling to refine distributions, training models on diverse datasets, and testing larger models to study how prompt design affects performance and attribute scores.

12 Conclusion

In this work, we have shown the correlation that exists between the similarity distribution of all query-passage pairs. This is further demonstrated by the increased performance typically observed when using more complex approaches, such as prompt engineering, as well as through human evaluation and LLM-based assessment. SynSim can serve as a valid solution that operates autonomously or can be integrated with classical methods, which require lower computational resources.

13 Ethical Considerations

This work studies the generation and evaluation of synthetic query–passage datasets using large language models (LLMs). While synthetic data enables scalable dataset creation and reduces the need for costly human annotation, it may also inherit biases present in the training data of the underlying models. In addition, LLM-generated queries can contain hallucinations or factual inaccuracies. In this study, such outputs are intentionally retained to evaluate their impact on similarity distributions, but they should be carefully filtered before real-world deployment.

We use only publicly available datasets (Natural Questions, HotpotQA, and FiQA) and do not collect new personal data.

14 Use of AI

Generative AI was utilized as a supportive tool for grammatical correction and stylistic editing to enhance the readability of this document. The core research, methodology, and final output remain the original work of the author, who has personally audited and approved every section for accuracy and consistency.

15 Limitation

This work explores future innovations while emphasising important limitations. Currently, the dataset is evaluated mainly through BERT-based similarity models and LLM-based judgements. Prompts are designed to provide product-related results, which are effective for GPT-based models but may not generalise to others. LLM assessments offer useful insights, yet the most reliable evaluation would involve expert human review of the entire dataset. Additionally, both human and LLM evaluations are limited to individual query-passage pairs rather than the full dataset, so the current results should be considered approximate.

References

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

Tiago Melo Almeida and Sérgio Matos. 2024. Exploring efficient zero-shot synthetic dataset generation for information retrieval. In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, pages 1214–1231. Association for Computational Linguistics.

Stefano Bistarelli and Marco Cuccarini. 2024. Feature selection on contextual embedding pushing the sparseness. In *AIxIA 2024 - Advances in Artificial Intelligence - XXIIIrd International Conference of the Italian Association for Artificial Intelligence, AIxIA 2024, Bolzano, Italy, November 25-28, 2024, Proceedings*, volume 15450 of *Lecture Notes in Computer Science*, pages 37–49. Springer.

Luiz Henrique Bonifacio, Hugo Queiroz Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Unsupervised dataset generation for information retrieval. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2387–2392. ACM.

Leonid Boytsov, Preksha Patel, Vivek Sourabh, Riddhi Nisar, Sayani Kundu, Ramya Ramanathan, and Eric Nyberg. 2024. Inpars-light: Cost-effective unsupervised training of efficient rankers. *Trans. Mach. Learn. Res.*, 2024.

Marco Braga, Pranav Kasela, Alessandro Raganato, and Gabriella Pasi. 2024. Synthetic data generation with large language models for personalized community question answering. In *IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2024, Bangkok, Thailand, December 9-12, 2024*, pages 360–366. IEEE.

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 *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, and 1 others. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 489–500. Association for Computational Linguistics.

Dmitrii Fomin, Philippe Lalanda, and Denis Morand. 2024. Automating the production of machine learning models for imbalanced datasets - application to industry 4.0. In *IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events, PerCom 2024 - Workshops, Biarritz, France, March 11-15, 2024*, pages 39–44. IEEE.

Johann Frei and Frank Kramer. 2023. Annotated dataset creation through large language models for non-english medical NLP. *J. Biomed. Informatics*, 145:104478.

Maliha Jahan, Helin Wang, Thomas Thebaud, Yinglun Sun, Giang Ha Le, Zsuzsanna Fagyal, Odette Scharenborg, Mark Hasegawa-Johnson, Laureano Moro-Velázquez, and Najim Dehak. 2024. Finding spoken identifications: Using GPT-4 annotation for an efficient and fast dataset creation pipeline. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 7296–7306. ELRA and ICCL.

Tristan Kenneweg, Philip Kenneweg, and Barbara Hammer. 2024. RAGVAL: automatic dataset creation

and evaluation for RAG systems. In *2nd International Conference on Foundation and Large Language Models, FLLM 2024, Dubai, United Arab Emirates, November 26-29, 2024*, pages 470–475. IEEE.

Alina Kramchaninova and Arne Defauw. 2022. Synthetic data generation for multilingual domain-adaptable question answering systems. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation, EAMT 2022, Ghent, Belgium, June 1-3, 2022*, pages 151–160. European Association for Machine Translation.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 957–966. JMLR.org.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Seongyun Lee, Hyunjae Kim, and Jaewoo Kang. 2023. LIQUID: A framework for list question answering dataset generation. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13014–13024. AAAI Press.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguistics*, 12:157–173.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pages 1941–1942. ACM.

Matthew Maufe, James Ravenscroft, Rob Procter, and Maria Liakata. 2022. A pipeline for generating, annotating and employing synthetic data for real world question answering. In *Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022 - System Demonstrations, Abu Dhabi, UAE, December 7-11, 2022*, pages 80–97. Association for Computational Linguistics.

Mohamed Moustafa, Amr Elrasad, Joseph Lemley, and Peter Corcoran. 2022. Dataset creation pipeline for camera-based heart rate estimation. In *Fifteenth International Conference on Machine Vision, ICMV 2022, Rome, Italy, November 18-20, 2022*, volume 12701 of *SPIE Proceedings*. SPIE.

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*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Will Orr and Kate Crawford. 2024. The social construction of datasets: On the practices, processes, and challenges of dataset creation for machine learning. *New Media Soc.*, 26(9):4955–4972.

Lucas Francisco Amaral Orosco Pellicer, Taynan Maier Ferreira, and Anna Helena Reali Costa. 2023. Data augmentation techniques in natural language processing. *Appl. Soft Comput.*, 132:109803.

Hossein A. Rahmani, Nick Craswell, Emine Yilmaz, Bhaskar Mitra, and Daniel Campos. 2024. Synthetic test collections for retrieval evaluation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 2647–2651. ACM.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A. Mann. 2021. Data augmentation can improve robustness. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 29935–29948.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Song Shuang and Muhammad Syafiq Bin Mohd Pozi. 2024. Using causal inference to solve uncertainty issues in dataset shift. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining, WSDM 2024, Merida, Mexico, March 4-8, 2024*, pages 1155–1157. ACM.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross J. Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nat.*, 631(8022):755–759.

Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023. Can chatgpt replace traditional KBQA models? an in-depth analysis of the question answering performance of the GPT LLM family. In *The Semantic Web - ISWC 2023 - 22nd International Semantic Web Conference, Athens, Greece, November 6-10, 2023, Proceedings, Part I*, volume 14265 of *Lecture Notes in Computer Science*, pages 348–367. Springer.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-llm collaborative annotation through effective verification of LLM labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024*, pages 303:1–303:21. ACM.

Vikas Yadav, Zheng Tang, and Vijay Srinivasan. 2024. PAG-LLM: paraphrase and aggregate with large language models for minimizing intent classification errors. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 2569–2573. ACM.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen R. McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking large language models for news summarization. *Trans. Assoc. Comput. Linguistics*, 12:39–57.

A Base Prompt Single-Hop

Here we provide the base prompt used for generating This prompt is used for **single-hop questions**:

```
You are a strict question generator.
Generate ONLY ONE Good Question from the given Document.
Output only the question. No explanations, no extra text, no punctuation

beyond the question itself.

Example 1:

Document: Several people have mentioned the obvious: inflation. But let's

assume we are talking about real
(inflation adjusted) prices. One argument is that populations keep rising

while the land does not change...

Good Question: What argument(s) support the claim that long-term housing

prices trend upward?

...

Predict:
Document:{}
Good Question:
```

—

B InPars Prompt Single-Hop

This prompt is used for **InPars single-hop question generation**, which emphasizes avoiding bad questions:

```
You are a strict question generator.
Generate ONLY ONE Good Question from the given Document.
Do NOT generate bad questions.
Output ONLY the Good Question.
Output only the question. No explanations, no extra text, no punctuation

beyond the question itself.

Example 1:

Document: Several people have mentioned the obvious: inflation. But let's

assume we are talking about real
(inflation adjusted) prices. One argument is that populations keep rising...

Bad Question: Why do house prices always go up?

Good Question: What argument(s) support the claim that long-term housing
```

```
prices trend upward?

...

Predict:
Document:{}
Good Question:
```

—

C InPars Prompt with Motivation Single-Hop

This variant also includes a **motivation for why bad questions are not ideal**:

```
You are a strict question generator.
For the given Document, generate:
ONLY ONE Good Question (answerable from the Document)
Motivation: a brief explanation of why the Bad Question is not ideal
Output MUST be structured exactly as one line Do NOT include anything else.

Example 1:

Document: Several people have mentioned the obvious: inflation. But let's

assume we are talking about real
(inflation adjusted) prices. One argument is that populations keep rising

while the land does not change...

Another argument is that the economy has a long-term upward trajectory.



Bad Question: Why do house prices always go up?

Motivation: Too general and misses key details about population growth,

income changes, and materials costs.

Good Question: What argument(s) support the claim that long-term housing

prices trend upward?

...

Predict:
Document:{}
Good Question:
```

—

D Prediction Samples

Below are sample predictions produced by dataset NQ with gemma3_12b

Index	Passage	Query	Base	Inpars	Motivation
1	In accounting, minority interest (or non-controlling interest) is the portion of a subsidiary corporation's stock that is not owned by the parent corporation.	What is non controlling interest on balance sheet	What is minority interest	What is the definition of minority interest	What percentage of a subsidiary can the parent company own
2	The fourth season of *Chicago Fire* premiered on October 13, 2015 and concluded on May 17, 2016. The season contained 23 episodes.	How many episodes are in Chicago Fire season 4	How many episodes are in season 4 of Chicago Fire	When did Chicago Fire season 4 begin	When did Chicago Fire season 4 premiere
3	"Love Will Keep Us Alive" was first performed by the Eagles in 1994 with lead vocals by Timothy B. Schmit.	Who sings Love Will Keep Us Alive by the Eagles	Who sang Love Will Keep Us Alive	Who sang Love Will Keep Us Alive	Who sings Love Will Keep Us Alive
4	Before being elected to federal office, Patrick Brown worked as a lawyer in Barrie.	Who is the leader of the Ontario PC Party	What did Brown do before being elected to office	What did he do before being elected to provincial office	What did Brown do before becoming an MPP
5	"Fishin' in the Dark" was released in June 1987 and certified Platinum by the RIAA in 2014.	Nitty Gritty Dirt Band Fishin' in the Dark album	What year was Fishin' in the Dark released	When was Fishin' in the Dark certified platinum	What is the significance of "Fishin' in the Dark"
6	Panning can be used to reduce stereo width, creating a dual mono signal when both channels are panned equally.	Converting stereo signal to mono signal is called	What does panning do	What does it mean to pan a stereo signal	What happens when you pan an audio mixer straight up

Table 5: Prediction samples generated by our models.