Generating Q&A Benchmarks for RAG Evaluation in Enterprise Settings

Simone Filice, Guy Horowitz, David Carmel
Zohar Karnin, Liane Lewin-Eytan, Yoelle Maarek
Technology Innovation Institute
Haifa, Israel
filice.simone@gmail.com, yoelle@yahoo.com
{guy.horowitz,david.carmel,zohar.karnin,liane.lewineytan}@tii.ae

Abstract

We introduce DataMorgana, a tool for generating synthetic Q&A benchmarks tailored to RAG applications in enterprise settings. DataMorgana enables customization of the generated benchmark according to the expected diverse traffic of the RAG application. It allows for specifying question types and their associated distribution via a lightweight configuration mechanism. We demonstrate via a series of quantitative and qualitative experiments that DataMorgana surpasses existing tools in terms of lexical, syntactic, and semantic diversity of the generated benchmark while maintaining high quality. We run our experiments over domain-specific and general-knowledge public datasets, as well as two private datasets from governmental RAG applications: one for citizens and the other for government employees. The private datasets have been shared with us by AI71, an AI company, which has integrated DataMorgana into its offerings. In addition, DataMorgana has been offered to about 150 researchers worldwide as part of the SIGIR'2025 LiveRAG Challenge held in Spring 2025.

1 Introduction

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Gao et al., 2023) has recently gained a great deal of popularity, especially in specialized domains. However, before adopting a RAG solution, it is critical to evaluate its effectiveness in the target environment, accounting not only for the environment's specific content (the RAG corpus) but also for its diverse types of users' needs.

Consider a corporate scenario where a company wishes to release a RAG-based question-answering experience over a private specialized corpus. In order to evaluate it, one requires a solid benchmark. In the absence of question or query logs, a common practice is to use an LLM to generate Q&A pairs from randomly selected documents within the corpus. The major risk in this approach is that the generated questions often have different properties than those that the experience owners envision. In particular, generated questions may not be realistic (e.g., contain too many technical terms), or diverse enough (e.g., not covering the many types of questions that could be asked by real users).

We propose a new approach to generate synthetic benchmarks with two key properties.
Lightweight and flexible customization: Configuring DataMorgana so that Q&A pairs are generated according to the expected traffic is done via free-text/natural language descriptions organized in categories, making customization accessible to non-AI specialists. This property is crucial because the system's user, typically a domain expert, should have control over the generation process as the one who truly understands the expected traffic.
Diverse generation: Our default setting offers a rich collection of orthogonal question category sets, which we denote as categorizations; these can be combined to create a combinatorial amount of question types, ensuring a diverse output. This results in an excellent starting point for the user, who can edit these categorizations, as modifying or removing question categories is far easier than creating them from scratch.

This approach is implemented in a tool called DataMorgana[1], which we describe in detail here. DataMorgana is fully deployed with AI71[2], an applied AI company focused on building intelligent agents that bring the benefits of AI to enterprise users in a responsible way. DataMorgana was made available to close to 150 researchers in the context of the SIGIR'2025 LiveRAG Challenge (Carmel et al., 2025). DataMorgana was used by competitors over March-May 2025 to train and test their RAG engine. It was also used by the Challenge organizers to generate the unseen synthetic test bench-

[1] This paper expands the work introduced by Filice et al. (2025).

[2] https://ai71.ai/

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 469–484
July 28-30, 2025 ©2025 Association for Computational Linguistics

marks on which the competitors were evaluated during a two-hour time window on May 20, 2025[3]. In this work, we focus on DataMorgana question generation capabilities, demonstrating through quantitative experiments that it achieves higher diversity than related tools or approaches without compromising quality. To keep our work focused, we leave for future work the analysis of the generated answers since this requires a completely different set of metrics, baselines, and ablations. Our key contributions are as follows.

1. We present DataMorgana, a synthetic benchmark generation tool with lightweight and flexible customization capabilities;

2. We propose a novel technique based on multi-question categorizations to support the generation of highly diverse benchmarks.

3. We conduct extensive experiments on both public and real-world proprietary datasets to assess how DataMorgana compares to existing benchmark generation methods in producing high-quality and more diverse questions across lexical, syntactic, and semantic dimensions.

2 Related Work

Recent advances in LLMs, with their tremendous zero-shot and few-shot generation capabilities, have led to many research efforts in creating synthetic test benchmarks for question answering (Fei et al., 2022; Dong et al., 2023; Yoon and Bak, 2023; Chen et al., 2024; Shakeri et al., 2020) and conversational dialog systems (Ling et al., 2020; Do et al., 2022). Ideally, an optimal test set would comprise a large set of real user questions from a query log, paired with “golden answers” provided by experts. In the absence of a perfect test set, we seek to generate questions similar to those asked by real users, along with answers inferred from a data source. A comprehensive taxonomy of generation approaches can be found in (Zhang et al., 2021; Long et al., 2024).

Generate then Filter: The common methodology for (question, answer) pair generation is to follow the *generate then filter* paradigm. Given a corpus of documents, select at first a subset of documents; then, for each document, leverage an LLM to generate some questions that can be answered by the given content. Next, ask the LLM to generate, for each of the questions, an answer or a set of answers based on the corresponding document. Finally, filter the generated (question, answer, document) tuples according to several criteria, such as semantic similarity with golden questions, diversity, and more (Yoon and Bak, 2023).

InPars (Jeronymo et al., 2023), Prompagator (Dai et al., 2022), and more recently ARES (Saad-Falcon et al., 2024), follow this paradigm. Via few-shot examples, an LLM is induced to generate relevant questions for a given document. Then, each (question, document) pair is scored and filtered according to their inner similarity, or if the associated document appears on top of the result list when the question is submitted as a query to a given IR system. Yuan et al. (2023) proposed a prompt-based approach to selecting high-quality questions while Shakeri et al. (2020) filtered the questions based on the generator’s perplexity score. Rackauckas et al. (2024) used real user queries as few-shot examples for synthetic query generation, to increase similarity with real traffic, and an LLM-as-a-judge approach for Q&A filtering.

Diversity: Uncontrolled generated content often tends to be monotonous and biased, hence limiting its applicability in downstream tasks (Long et al., 2024). The diversity of generated data is crucial for generating synthetic samples that mimic the diversified nature of real-world data, thereby preventing over-fitting and bias during model training or evaluation. Yoon and Bak (2023) improve question diversity by training the model to generate a new question that differs from previously generated questions. Eo et al. (2023) enhance diversity by training the generator to cover various types of questions per document, based on interrogative question words ({Who, When, What, Where, Why, How}).

Control: Recent studies have suggested enhancing administrative control over the types of generated questions. *Know Your RAG* (de Lima et al., 2024) is a system designed to generate questions from a predefined taxonomy of question types. The generator decomposes the document into statements, and then, depending on the question type, a single statement or an aggregation of multiple statements is used as a basis for question generation.

RAGAs (RAGAS, 2025) is a popular evaluation

[3] More details about the Challenge can be found at `https://liverag.tii.ae`.

tool for RAG systems that supports the generation of a synthetic Q&A benchmark. Similarly to *Know Your RAG*, RAGAs considers a predefined set of different question types (single-hop vs multi-hop, specific vs abstract), as well as the user persona (senior, junior, etc.). This enriches the type of generated questions and improves diversity. DeepEval (DeepEval, 2025) generates a synthetic Q&A benchmark, while encouraging diversity by an evolutionary process where new questions are generated according to pre-defined evolution rules.

In contrast, DataMorgana, which we describe next, controls the diversity level with finer granularity via a configuration mechanism. We then discuss how DataMorgana compares to some of the approaches discussed above in terms of diversity.

3 DataMorgana System Description

DataMorgana is designed to generate synthetic benchmarks for training and testing primarily RAG (and possibly other) systems that offer question-answering capabilities. It operates in two stages: a configuration stage, during which the DataMorgana admin user specifies their needs, and a generation stage, during which DataMorgana leverages the input configuration to generate, with the assistance of an LLM, the desired benchmark.

3.1 Configuration Stage

In the configuration stage, the user defines question categorizations. Each categorization consists of one or more mutually exclusive categories; thus, a generated question can belong to only one category from each categorization.

The configuration is done either through a JSON file or a visual interface. See Appendix A, Figure 2 for an example of the JSON format. The example lists categorizations *Question factuality* with categories of *factoid* or *open-ended*, *Phrasing* with categories describing whether the question is concise or verbose and whether it is naturally formed, or phrased as a search query. The question categorization can also be leveraged to specify the type of users who would issue the question. For example, a categorization could be *End-user expertise*, indicating the user's familiarity with the documents' content, with categories of *novice* or *expert*, as shown in Appendix A, Figure 3. In a healthcare RAG application, one could add *patient, doctor,* and *public health authority* as categories under a RAG system user categorization. Additional examples can be found in Table 5 in the Appendix, containing general-purpose question categorizations and their respective categories, which can be used for most corpora.

The user may specify, in addition to the categories of a categorization, a probability distribution over them, to determine their frequency in the generated dataset. Formally, the categorization definition results in k categorizations $C_1, \ldots, C_k$, each consisting of categories $\{c_{i,1}, \ldots, c_{i,n_i}\}$ and corresponding probabilities $\{p_{i,1}, \ldots, p_{i,n_i}\}$.

3.2 Generation Stage

The benchmark is built incrementally one Q&A pair (q_i, a_i) at a time, via the following procedure

- A document d_i is sampled from the corpus.
- For each categorization $C_1, \ldots, C_k$, we sample a single category $c_{i,j} \in C_j$ according to the distribution provided with C_j.
- A prompt (see example in Figure 4 of the Appendix) is built asking the LLM to generate a question based on document d_i belonging to the categories $c_{i,1}, \ldots, c_{i,k}$. The chosen LLM is then invoked to generate the question.

Note that this methodology is simple and lightweight by design, allowing for quick development iterations. We intentionally try to avoid approaches with a costly pre-processing stage (e.g., building a knowledge graph (RAGAS, 2025), performing heavy analysis on the document (de Lima et al., 2024)) or multiple invocations for post-processing (e.g., evolving a question (DeepEval, 2025)).

4 Baselines & Corpora

Baselines. We compare DataMorgana with the following synthetic data generation methods:

Vanilla is a strategy that repeatedly uses the same exact process to generate questions from different documents, namely, the LLM instructions appearing in the prompt are always the same, and the only part that varies is the input document. This is probably the most common, albeit straightforward, strategy to generate synthetic benchmarks (Chen et al., 2024; Wang et al., 2024a,b; Li et al., 2024).

Know Your RAG. We re-implemented the solution proposed by de Lima et al. (2024) described in

Section 2. The original solution generates four question types: single-fact, reasoning, summary, and unanswerable questions. We excluded the latter since, while it is fitting for reading comprehension, it is too challenging in a RAG context to guarantee that no document in the corpus can answer the question, making it difficult to assess an answer to the question. More importantly, allowing unanswerable questions introduces unbounded freedom in diversity, which contradicts our focus on measuring and analyzing diversity in this paper. Including them would create a setup that lacks a meaningful basis for comparison. We, in fact, verify as a quality test that a question is answerable.

DeepEval. We chose DeepEval (DeepEval, 2025) as a representative of unpublished commercial solutions. It is well adopted (as of May 2025, their git repo has 6.4K stars and 567 forks), and their data generation code is easy to run and flexible enough to allow generating multiple questions per document. We used their default setting that enables evolving questions with one evolution step, where the type is drawn uniformly at random from seven possible evolutions.

For a fair comparison, all tested generation methods leverage `Claude-3.5 Sonnet v2`[4] with default parameters as the LLM backbone [5].

Corpora. To showcase DataMorgana's capabilities, we generated synthetic data from four corpora. The first two are public datasets: Wikipedia (from 2018) and the CORD-19 Open Research Dataset (CORD-19) (Wang et al., 2020). Further details about these corpora are in Appendix B. The other two are domain-specific proprietary datasets shared with us by an AI company[2] that offers RAG-based chatbots to governmental entities. One of these, referred to as GovExternal, is used to allow citizens to ask about processes related to a governmental agency. The corpus is derived from semi-structured proprietary material, containing mostly natural language text with some embedded small tables, represented as text via markdown. The other one, denoted as GovInternal, is for internal government employees. The corpus contains internal reports containing structured partitions such as sections, subsections, and some bullet points. These are encoded in the text via markdown. The intended use is for employees to be able to ask a chatbot about the content of the reports. In both cases, the corpus is relatively small, covering no more than a few hundred documents.

5 Example Use Case

In this section, we demonstrate the process of using DataMorgana to generate a synthetic dataset over a corpus. For this demonstration, we chose the COVID-QA benchmark (Möller et al., 2020), which contains a collection of questions that were manually composed by field experts, based on documents in the CORD-19 corpus. It is a public corpus, allowing us to discuss its characteristics, yet it exhibits many similarities with proprietary ones. It covers topics typically not included at this level of detail in Wikipedia or other general-knowledge corpora and features highly specialized content with domain-specific jargon.

A user of DataMorgana has a good understanding of the types of questions that could be asked over the corpus and would like to configure the system to generate such questions. We use a small sample of questions from Covid-QA (see human-generated questions in Table 2) to represent the type of questions the user aims to create.

Consider first a scenario where the user applies the Vanilla approach. Table 2 contains sample questions generated in this scenario. One can see that the LLM is inherently biased in producing questions that are longer, contain many details, and tend to be about a specific topic rather than open-ended. They lack fidelity in that many of them are too different from the human-generated questions. It is likely possible to modify the prompt to avoid these unrealistic questions. However, diversity will be much tougher to solve. The resulting set is likely to cover only a small portion of the sought distribution, e.g., by containing only factoid-seeking questions, as opposed to open-ended questions, or vice versa.

With DataMorgana, the user can define categorizations matching the target questions, and cover the different question types. In this case, the user defines the three categorizations shown in Table 1.

Table 2 contains sample questions created by DataMorgana using this configuration. It is easy to see that the questions are, on one hand, of the same nature as the human-generated questions (fidelity),

[4] https://www.anthropic.com/claude/sonnet

[5] The LLM prompts we used within DataMorgana are not specifically engineered for Claude-3.5 Sonnet v2, but are expected to work well with other similar-size LLMs.

Categorization	Category	Description
Factuality	factoid	question seeking a specific, concise piece of information or a short fact about a particular subject, such as a name, date, or number (e.g., *'When was Napoleon born?'*).
	open-ended	question inviting detailed or exploratory responses, encouraging discussion or elaboration. (e.g., *'what caused the French revolution?')*.
Phrasing	concise-and-natural	phrased in the way people typically speak, reflecting everyday language use, without formal or artificial structure. It is a concise, direct question consisting of less than 10 words (e.g., *'what's the weather like in Paris now?'*).
End-user expertise	clinical researcher	A clinical researcher who uses the system to access population health data, conduct initial patient surveys, track disease progression patterns, etc.

Table 1: Configuration used to mimic the Covid-QA dataset.

in that they are short and discuss technical topics, and on the other, are sufficiently diverse, e.g., contain both factoid and open-ended questions. More examples generated by the Vanilla method, as well as additional examples from DeepEval and Know Your RAG systems, can be found in Appendix C.

Random Sample of Questions generated by DataMorgana
Is COVID more infectious than MERS?
How do calcium inhibitors block flavivirus infections?
How deadly was COVID compared to SARS and MERS?
How effective are neutralizing antibodies in fighting hepatitis C?
What age groups are most vulnerable to seasonal flu complications?
What factors increase risk of hantavirus outbreaks?
When do RSV infections peak in children?
What were the main symptoms of early COVID-19 cases?
Random Sample of Questions generated by Vanilla
What were the main routes of transmission for SARS-CoV-2 in the early stage of the outbreak in Wuhan, and which one was more significant?
How do humans typically get infected with hantavirus, and what activities put people at higher risk of infection?
How common are co-infections in people who have influenza, and why is this important for treatment?
Random Sample of human-generated Questions
How does MARS-COV differ from SARS-COV?
How was HFRS first brought to the attention of western medicine ?
What can respiratory viruses cause?
What is MERS mostly known as?
What is RANBP2?
What is the transmission of MERS-CoV is defined as?
What reduces the antimicrobial activities of alveolar macrophages?
Where did SARS-CoV-2 originate?

Table 2: Random sample of questions generated by DataMorgana, Vanilla, and humans for qualitative comparison.

6 Quantitative Study

As per (Alaa et al., 2022), three key aspects should be considered to assess the quality of synthetic data quality: generalization, fidelity, and diversity. Generalization applies to models like GANs that generate data based on a training set of real examples, making it irrelevant to our setting. In our context, fidelity translates to the quality of individual questions, namely the extent to which generated questions represent a plausible way a real user could interact with the system, while diversity ensures that the generated questions cover all or at least many of the questions asked by humans.

As discussed before, while achieving fidelity is essential, it can be achieved with simple methods. Diversity, however, which is no less significant, is more challenging. In the rest of this section, we first discuss quality and then diversity, after establishing that quality is maintained.

6.1 Measuring Quality

Following the definitions of Fu et al. (2024), we consider three metrics related to the question text quality: Fluency (Oh et al., 2023), Clarity (Ousidhoum et al., 2022), Conciseness (Cheng et al., 2021), and 3 metrics assessing the match between the question and document used to generate it: Relevance (Oh et al., 2023), Consistency (Honovich et al., 2022), Answerability (Ghanem et al., 2022) (see Appendix D for formal definitions). We generated 200 questions for each (corpus, method) pair. We computed all six metrics for each of these 16 benchmarks using a strong LLM (Claude Sonnet 3.5). Appendix D provides further details about the metrics, as well as the LLM prompts we used.

Results, reported in Table 3, show a consistent pattern along the benchmarks and metrics. We see near-perfect results for text quality and good results for passage match. Moreover, we see that DataMorgana is on par or better than the baselines, with the exception of passage match when compared with Vanilla, likely due to Vanilla not being constrained by the type of questions it should generate.

Next, we present an analysis of diversity and coverage across methods.

6.2 Measuring Diversity

To estimate the diversity of the generated benchmark, we use the following metrics, as suggested in (Shaib et al., 2024): to capture lexical diversity, we

Corpus	Generation Method	Text Quality			Passage Match		
		Fluency	Clarity	Conciseness	Relevance	Consistency	Answerability
Cord-19	DataMorgana	1.00	1.00	1.00	0.97	0.97	0.89
Cord-19	DeepEval	1.00	0.99	1.00	0.94	0.92	0.82
Cord-19	Know Your RAG	1.00	1.00	1.00	1.00	0.99	0.99
Cord-19	Vanilla	1.00	1.00	0.99	0.99	0.99	0.98
GovInternal	DataMorgana	0.99	1.00	1.00	0.98	0.99	0.95
GovInternal	DeepEval	0.96	0.94	0.98	0.88	0.84	0.75
GovInternal	Know Your RAG	0.99	0.99	1.00	1.00	0.99	0.98
GovInternal	Vanilla	1.00	1.00	1.00	1.00	1.00	1.00
GovExternal	DataMorgana	0.93	0.93	1.00	0.96	0.91	0.82
GovExternal	DeepEval	0.99	0.95	0.94	0.83	0.82	0.60
GovExternal	Know Your RAG	0.99	0.92	0.95	0.92	0.86	0.81
GovExternal	Vanilla	0.90	1.00	1.00	0.98	0.97	0.96
Wiki	DataMorgana	1.00	0.99	0.99	0.97	0.96	0.86
Wiki	DeepEval	1.00	0.96	0.94	0.84	0.74	0.56
Wiki	Know Your RAG	1.00	0.98	0.99	0.94	0.84	0.79
Wiki	Vanilla	1.00	1.00	1.00	0.99	0.97	0.96

Table 3: Quality results on all benchmarks and generation methods.

use N-Gram Diversity (NGD), which measures the fraction of unique n-grams (with $n \in [1, 4]$), Self-Repetition score (SR), which counts the number of instances containing at least one n-gram ($n = 4$) appearing elsewhere, and Word Compression Ratio (word-CR), which measures the compression ratio of the file containing the generated questions. To capture syntactic diversity, we measure part-of-speech compression (PoS-CR), where we convert sentences to their PoS tags and compute the compression ratio of the resulting file. To capture semantic diversity, we compute the Homogenization Score (HS), which represents the average similarity between all question embeddings. We provide the full formal definitions of these metrics in Appendix E.1.

We evaluated all four methods over the four corpora by generating 1K-3K questions per corpus (3K for Wiki, 2K for Cord-19 and GovExternal, and 1K for GovInternal), and measuring their diversity via those metrics. Table 4 contains the metrics computed for the different benchmarks. First, we note that, as expected, Vanilla shows inferior diversity across all metrics. For both semantic and syntactic diversity, one can see a clear advantage of DataMorgana compared to the baselines. For lexical diversity DataMorgana surpasses Know Your Rag, and has close performance to DeepEval in the word-CR and NGD metrics. Upon inspecting DeepEval's questions (See Table 6 in the Appendix for some examples), we noticed they tend to be verbose and use multiple technical terms. Such a property increases lexical diversity, but does not contribute to semantic or syntactic diversity, which aligns with the experiment results.

Corpus	Model		Lex		Syn	Sem
		NGD(↑)	SR	w-CR	P-CR	e-HS
GovExt	VL	1.122	0.982	6.842	10.734	0.260
	KYR	1.777	0.833	4.910	7.292	0.269
	DE	1.787	0.865	4.643	7.710	0.232
	DM	**1.838**	**0.670**	**4.415**	**6.642**	**0.178**
GovInt	VL	1.248	0.992	6.552	9.484	0.396
	KYR	2.274	0.747	4.256	6.515	0.315
	DE	**2.682**	0.507	**3.466**	5.866	0.269
	DM	2.469	**0.482**	3.802	**5.795**	**0.213**
Cord-19	VL	1.517	0.920	5.576	7.861	0.301
	KYR	2.358	0.613	3.879	6.271	0.265
	DE	2.415	0.644	**3.535**	5.885	0.251
	DM	**2.536**	**0.372**	3.701	**5.583**	**0.249**
Wiki	VL	2.662	0.533	2.665	5.824	0.068
	KYR	2.981	0.144	2.488	5.864	0.074
	DE	2.879	0.371	**2.477**	5.631	0.067
	DM	**3.016**	**0.140**	2.502	**5.397**	**0.052**

Table 4: Diversity scores of different synthetic datasets. In bold, the best results, underlined the results whose difference w.r.t. the best result is not statistically significant (see Appendix E.2 for details). We use the following shorthands for models, VL: Vanilla, KYR: Know Your RAG, DE: DeepEval, DM: DataMorgana, and for metrics, w-CR: word-CR, P-CR: part-of-speech-CR, e-HS: embedding-HS. For all metrics other than NGD, lower is better.

In addition, we measured the diversity of the methods in a setting where the number of documents is limited. Here, multiple questions must be generated from a single document, potentially limiting the diversity of the questions. Figure 1 reports the results for the four explored corpora. We set to 200 the total number of generated questions and increased the number of documents used to generate them, from 20 (i.e., 10 questions per document) to 147 for Cord-19 and 200 for GovExternal. For each of the generated benchmarks, corresponding to a different number of documents (X-axis), we measured the PoS-CR metric (Y-axis). We see that DataMorgana consistently outperforms

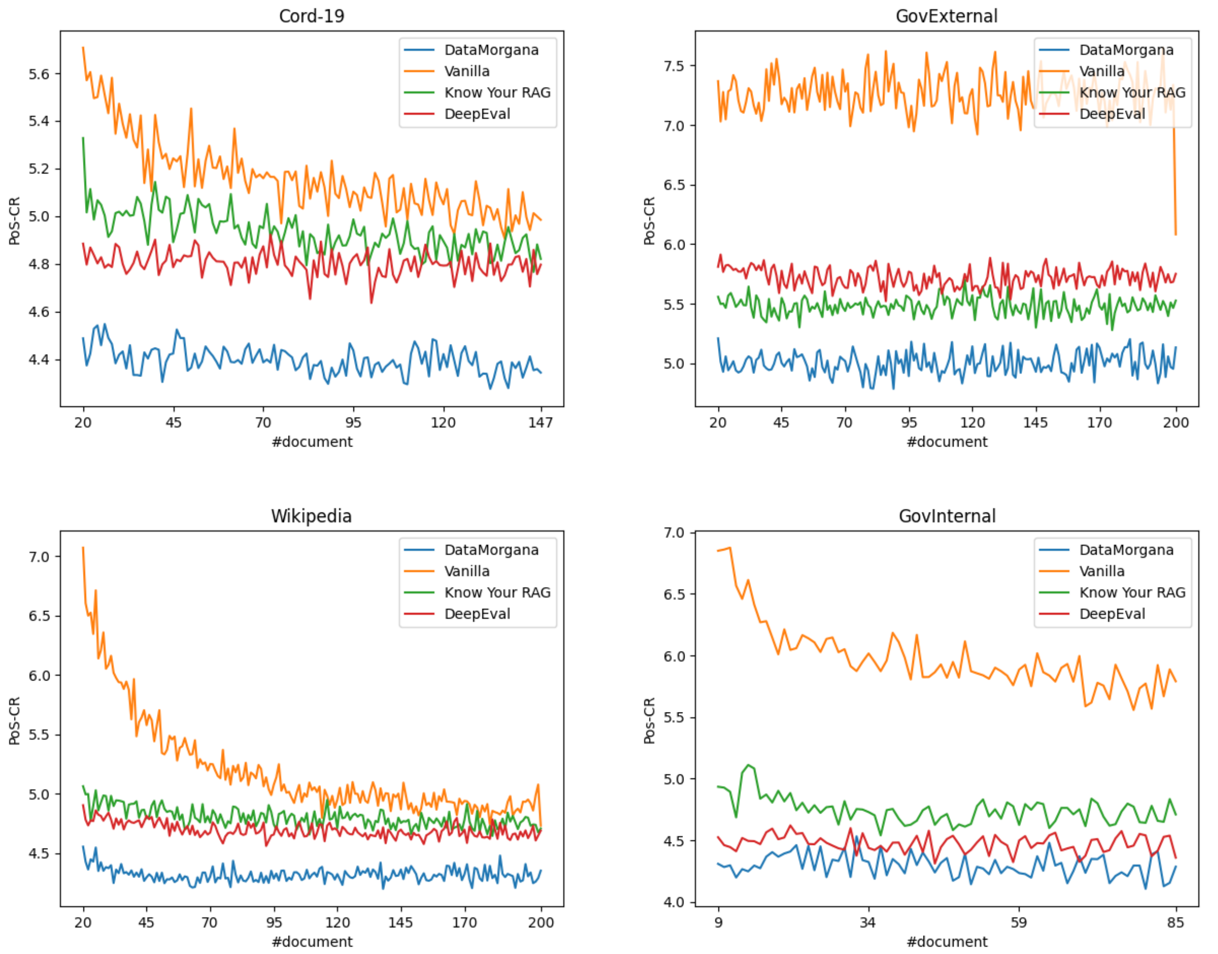

Figure 1: Syntactic PoS-CR(↓) diversity of synthetic benchmarks containing 200 questions generated from an increasing number of documents, over the CORD-19 (top-left), Wikipedia (bottom-left), GovExternal (top-right), and GovInternal (bottom-right) corpora.

the baselines for all settings.

7 Conclusion

We presented DataMorgana, a benchmark generation tool that offers simple, yet rich configuration capabilities, to tailor synthetic benchmarks to the expected traffic of a RAG application.

Through both qualitative and quantitative analyses, we showed that DataMorgana generates questions that are at the same level of quality, yet are significantly more diverse than those produced by other question generation tools. These tools typically either leave the choice of question type to the LLM, or use internal mechanisms to control question diversity.

While DataMorgana was originally designed for RAG systems evaluation, it is generic enough to be used to evaluate any Q&A system. We intend to introduce soon additional capabilities for generating other types of benchmarks, such as conversations. Additionally, we plan to extend our study of diversity, for example, to make sure that for long documents containing multiple topics, we generate questions covering all contained topics.

DataMorgana has been deployed at AI71 and has been extensively used by close to 150 researchers as part of the SIGIR'2025 LiveRAG Challenge over March-May 2025.

Acknowledgments

We are, as always grateful, to our awesome colleagues and partners at TII (Tom Beer, Ran Tavory and Oren Somekh) and AI71 (Darshan Agarwal, Mehdi Ghissassi, and Ramy Makary) for their insights and support and hard work deploying DataMorgana.

References

Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela van der Schaar. 2022. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*, pages 290–306. PMLR.

Valeriia Bolotova, Vladislav Blinov, Falk Scholer, W. Bruce Croft, and Mark Sanderson. 2022. A non-factoid question-answering taxonomy. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 1196–1207, New York, NY, USA. Association for Computing Machinery.

David Carmel, Simone Filice, Guy Horowitz, Yoelle Maarek, Oren Somekh, and Ran Tavory. 2025. The liverag challenge at sigir 2025. SIGIR '25, New York, NY, USA. Association for Computing Machinery.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Yi Cheng, Siyao Li, Bang Liu, Ruihui Zhao, Sujian Li, Chenghua Lin, and Yefeng Zheng. 2021. Guiding the growth: Difficulty-controllable question generation through step-by-step rewriting. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5968–5978.

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*.

Rafael Teixeira de Lima, Shubham Gupta, Cesar Berrospi, Lokesh Mishra, Michele Dolfi, Peter Staar, and Panagiotis Vagenas. 2024. Know your rag: Dataset taxonomy and generation strategies for evaluating rag systems. *arXiv preprint arXiv:2411.19710*.

DeepEval. 2025. DeepEval: Synthesizers.

Xuan Long Do, Bowei Zou, Liangming Pan, Nancy Chen, Shafiq Joty, and Aiti Aw. 2022. Cohs-cqg: Context and history selection for conversational question generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 580–591.

Chenhe Dong, Ying Shen, Shiyang Lin, Zhenzhou Lin, and Yang Deng. 2023. A unified framework for contextual and factoid question generation. *IEEE Transactions on Knowledge and Data Engineering*, 36(1):21–34.

Sugyeong Eo, Hyeonseok Moon, Jinsung Kim, Yuna Hur, Jeongwook Kim, SongEun Lee, Changwoo Chun, Sungsoo Park, and Heuiseok Lim. 2023. Towards diverse and effective question-answer pair generation from children storybooks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6100–6115, Toronto, Canada. Association for Computational Linguistics.

Zichu Fei, Qi Zhang, Tao Gui, Di Liang, Sirui Wang, Wei Wu, and Xuan-Jing Huang. 2022. Cqg: A simple and effective controlled generation framework for multi-hop question generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6896–6906.

Simone Filice, Guy Horowitz, David Carmel, Zohar Karnin, Liane Lewin-Eytan, and Yoelle Maarek. 2025. Generating diverse Q&A benchmarks for RAG evaluation with DataMorgana.

Weiping Fu, Bifan Wei, Jianxiang Hu, Zhongmin Cai, and Jun Liu. 2024. Qgeval: Benchmarking multi-dimensional evaluation for question generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11783–11803.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Bilal Ghanem, Lauren Lutz Coleman, Julia Rivard Dexter, Spencer von der Ohe, and Alona Fyshe. 2022. Question generation for reading comprehension assessment by modeling how and what to ask. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2131–2146.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. True: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920.

Vitor Jeronymo, Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, Jakub Zavrel, and Rodrigo Nogueira. 2023. Inpars-v2: Large language models as efficient dataset generators for information retrieval. *arXiv preprint arXiv:2301.01820*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur 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. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Jiarui Li, Ye Yuan, and Zehua Zhang. 2024. Enhancing llm factual accuracy with rag to counter hallucinations: A case study on domain-specific queries in private knowledge-bases. *arXiv preprint arXiv:2403.10446*.

Yanxiang Ling, Fei Cai, Honghui Chen, and Maarten de Rijke. 2020. Leveraging context for neural question generation in open-domain dialogue systems. In *Proceedings of The Web Conference 2020*, pages 2486–2492.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11065–11082.

Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. COVID-QA: A question answering dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Shinhyeok Oh, Hyojun Go, Hyeongdon Moon, Yunsung Lee, Myeongho Jeong, Hyun Seung Lee, and Seungtaek Choi. 2023. Evaluation of question generation needs more references. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6358–6367.

Nedjma Ousidhoum, Zhangdie Yuan, and Andreas Vlachos. 2022. Varifocal question generation for fact-checking. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2532–2544.

Zackary Rackauckas, Arthur Câmara, and Jakub Zavrel. 2024. Evaluating rag-fusion with ragelo: an automated elo-based framework. *arXiv preprint arXiv:2406.14783*.

RAGAS. 2025. Ragas: Testset generation for RAG.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. Ares: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354.

Chantal Shaib, Joe Barrow, Jiuding Sun, Alexa F. Siu, Byron C. Wallace, and Ani Nenkova. 2024. Standardizing the measurement of text diversity: A tool and a comparative analysis of scores.

Siamak Shakeri, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Feng Nan, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. End-to-end synthetic data generation for domain adaptation of question answering systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5445–5460, Online. Association for Computational Linguistics.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Shuting Wang, Jiongnan Liu, Shiren Song, Jiehan Cheng, Yuqi Fu, Peidong Guo, Kun Fang, Yutao Zhu, and Zhicheng Dou. 2024a. Domainrag: A chinese benchmark for evaluating domain-specific retrieval-augmented generation. *arXiv preprint arXiv:2406.05654*.

Shuting Wang, Jiejun Tan, Zhicheng Dou, and Ji-Rong Wen. 2024b. Omnieval: An omnidirectional and automatic rag evaluation benchmark in financial domain. *arXiv preprint arXiv:2412.13018*.

Hokeun Yoon and JinYeong Bak. 2023. Diversity enhanced narrative question generation for storybooks. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Xingdi Yuan, Tong Wang, Yen-Hsiang Wang, Emery Fine, Rania Abdelghani, Hélène Sauzéon, and Pierre-Yves Oudeyer. 2023. Selecting better samples from pre-trained llms: A case study on question generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12952–12965.

Ruqing Zhang, Jiafeng Guo, Lu Chen, Yixing Fan, and Xueqi Cheng. 2021. A review on question generation from natural language text. *ACM Transactions on Information Systems (TOIS)*, 40(1):1–43.

A DataMorgana Configuration

Figures 2 and 3 provide some illustrative snippets of the JSON configuration file for question categorizations. Table 5 details a set of general-purpose question categorizations and their respective categories, which can be used for most corpora. Figure 4 contains the prompt used by DataMorgana for generated a Q&A pair.

```
{
    "categorization": {
        "categorization_name": "question-factuality",
       "description": "This categorization distinguishes between factoid and non-factoid questions.",
        "categories": [
            {
                "name": "factoid",
                "probability": 0.25,
               "description": "a question seeking a specific, concise piece of information or a short
                ↪  fact about a particular subject, such as a name, date, or number."
            },
            {
                "name": "non-factoid-experience",
                "probability": 0.75,
                "description": "A question to get advice or recommendations on a particular topic."
            }
        ]
    }
}
```

Figure 2: Example of Question Categorization including factoid and non-factoid "experience" questions, as defined in the six types (i.e., instructions, reason, evidence-based, comparison, experience, and debate) of non-factoid questions suggested in (Bolotova et al., 2022).

```
{
   "categorization": {
        "categorization_name": "user-expertise-categorization",
        "description": "This categorization defines the level of expertise of end-users.",
        "categories": [
            {
                "name": "expert",
                "probability": 0.5,
                "description": "a specialized user with deep understanding of the corpus."
            },
            {
                "name": "novice",
                "probability": 0.5,
                "description": "a regular user with no understanding of specialized terms."
            }
        ]
    }
}
```

Figure 3: Example of question categorization relating to the user who might express it.

```
You are a user simulator that should generate a question to start a conversation.

The question must be about facts discussed in the document you will now receive.
Return only the question and its answer without any preamble.
Write the question-answer pair in the following JSON format:
{”question”: <question>, ”answer”: <answer>}.

### The generated question should be about facts from the following document:
[document (d_i)]

### The generated question must reflect a user with
the following characteristics:
    - [description of user category 1 (u_1)]
    - [description of user category 2 (u_2)]
    . . .

NOTE: you must use this information only when generating the question.
Instead, while answering the question you must ignore all the user characteristics.

### The generated question must have the following characteristics:
    - The question must be understandable by a reader who does    not have access to the document
    and does not even know what the document is about.
    Therefore, never refer to the author of the document or the document itself.
    - The question must include all context needed for comprehension.
    - The question must be answerable using solely the information presented in the document.
    - [description of question category 1 (c_1)]
    - [description of question category 2 (c_2)]
    . . .

### The answer to the generated question must have the following characteristics:
    - It must be very similar to the document in terms of terminology and phrasing.
    - It should only contain claims that directly appear in the document
    or that are directly deducible from it.
    - It must be understandable by a reader who does not have access to the document.
    Therefore, never refer to the author of the document or the document itself.
    - It must not assume or contain any information about the user,
    unless it is explicitly revealed in the question.
```

Figure 4: QA generation Prompt Template.

Categorization	Category	Description
Factuality	factoid	question seeking a specific, concise piece of information or a short fact about a particular subject, such as a name, date, or number (e.g., *'When was Napoleon born?'*).
	open-ended	question inviting detailed or exploratory responses, encouraging discussion or elaboration. (e.g., *'what caused the French revolution?').*
Premise	direct	question that does not contain any premise or any information about the user) (e.g., *'what is the fee for speeding in Italy?'*)
	with-premise	question starting with a very short premise, where the user reveals their needs or some information about himself (e.g., *'I have an H1-B visa for the United States. Is there a limit to how many times I can exit and enter the country in a year?'*).
Phrasing	concise-and-natural	phrased in the way people typically speak, reflecting everyday language use, without formal or artificial structure. It is a concise, direct question consisting of less than 10 words (e.g., *'what's the weather like in Paris now?'*).
	verbose-and-natural	phrased in the way people typically speak, reflecting everyday language use, without formal or artificial structure. It is a relatively long question consisting of more than 9 words (e.g., *'I thought of visiting Paris this year, not sure when is the best time. How is it like in the summer?'*).
	short-search-query	phrased as a typed web query for search engines (only keywords, without punctuation and without a natural-sounding structure). It consists of less than 7 words (e.g., *'Paris weather August'*).
	long-search-query	phrased as a typed web query for search engines (only keywords, without punctuation and without a natural-sounding structure). It consists of more than 6 words (e.g., *'Paris, France temperature humidity climate summer vs fall'*).
Linguistic variation	similar-to-document	phrased using the same terminology and phrases appearing in the document (e.g., for the document 'The Amazon River has an average discharge of about 215,000–230,000 m3/s', *'what is the average discharge of the Amazon river'*).
	distant-from-document	phrased using terms completely different from the ones appearing in the document (e.g., for a document 'The Amazon River has an average discharge of about 215,000–230,000 m3/s', *'How much water run through the Amazon?'*).
User expertise	expert	The user asking the question is a specialized user with a deep understanding of the corpus.
	novice	The user asking the question is a regular user with no understanding of specialized terms.

Table 5: Default configuration for DataMorgana. The examples in parentheses are for illustration only and are not necessarily part of the description to be used for generation.

B Experimental Setup Details

Corpora:

- COVID-19 Open Research Dataset (**CORD-19**) (de Lima et al., 2024) contains scientific papers on COVID-19 and related historical coronavirus research. We selected the 147 articles that biomedical experts used when generating the questions appearing in the COVID-QA dataset (Möller et al., 2020).

- **Wikipedia** is a free online encyclopedia that contains millions of articles about general human knowledge. We considered the 2682 articles containing answers to the questions in the test set of the NQ dataset (Kwiatkowski et al., 2019), containing questions asked by real users when using a search engine.

C Case Study

In Table 6, we report a random set of questions about different articles from the CORD-19 corpus, generated by different methods.

D Details of Quality experiments

We used an LLM-as-a-judge strategy for assessing quality. The chosen LLM is Claude-3.5-sonnet, with the prompt given in Figures 5,6,7). The prompt contains the definitions of the metrics, as well as few-shot examples. For each question and metric, the LLM provides a score between 1 and 3. We average the scores over each benchmark and normalize this mean by shifting it to be in the range of $[0, 1]$ via a linear shift $(\text{mean} - 1)/2)$. Namely, the presented scores are near perfect when reaching 1 and have a lower quality as they reach 0.

Model	Random Sample of Questions
Vanilla	How common are co-infections in people who have influenza, and why is this important for treatment?
	How do humans typically get infected with hantavirus, and what activities put people at higher risk of infection?
	How do humans typically get infected with pathogenic arenaviruses?
	How does the protein Prohibitin (PHB) affect the life cycle of the lymphocytic choriomeningitis virus?
	What are the main approaches being explored for developing a universal influenza vaccine using viral vectors?
	What are the main clinical symptoms and warning signs of severe adenovirus type 55 infection in otherwise healthy adults?
	What specific protective equipment and safety measures were required for healthcare workers conducting CT scans of COVID-19 patients?
	What were the main routes of transmission for SARS-CoV-2 in the early stage of the outbreak in Wuhan, and which one was more significant?
Know Your RAG	By how much did pneumonia deaths in children decrease between 2000-2013 due to new vaccines?
	How do virus-vectored flu vaccines compare to traditional vaccines in terms of safety and immune response?
	How does 2-bromopalmitic acid affect hantavirus host cell mineralization patterns?
	How does EGR1 deficiency affect BIRC5 expression during VEEV infection?
	What genetic similarities does the French BCoV strain share with Asian coronavirus strains?
	What safer alternative to live virus can be used for arenavirus neutralization testing?
	What starting material did the engineered E. coli platform use to generate glucose-1-phosphate for UDP-sugar synthesis?
	What was Germany's COVID-19 infection rate compared to other European countries during early pandemic interventions in March 2020?
DeepEval	How did World War 1's social and economic conditions make the Spanish flu pandemic more deadly, leading to over 20 million deaths?
	How do environmental factors like habitat fragmentation, and climate patterns affect hantavirus outbreaks and rodent populations in the Americas?
	How do respiratory viruses affect the airways?
	How would Australian-Japanese biomedical research collaboration be different today if the AIFII and ConBio conferences had never taken place?
	How would scientists use VP1 sequencing and viral testing to identify meningitis infections if an outbreak happened today?
	What are the average and highest percentage increases in COVID-19 cases predicted for China by FPASSA-ANFIS?
	What are the advantages and challenges of using Ad5 as a vaccine vector, particularly regarding stability, storage, delivery, and immunity issues?
	What's the difference between TIV, QIV, and LAIV flu vaccines, and which one provides the best protection?
	Which caspases are activated, and at what concentrations, when HT-29 cells are treated with Cu2 compared to untreated cells?
DataMorgana	Is COVID more infectious than MERS?
	How do calcium inhibitors block flavivirus infections?
	How deadly was COVID compared to SARS and MERS?
	How effective are neutralizing antibodies in fighting hepatitis C?
	What age groups are most vulnerable to seasonal flu complications?
	What factors increase risk of hantavirus outbreaks?
	When do RSV infections peak in children?
	What were the main symptoms of early COVID-19 cases?
Humans	How does MARS-COV differ from SARS-COV?
	How was HFRS first brought to the attention of western medicine ?
	What can respiratory viruses cause?
	What is MERS mostly known as?
	What is RANBP2?
	What is the transmission of MERS-CoV is defined as?
	What reduces the antimicrobial activities of alveolar macrophages?
	Where did SARS-CoV-2 originate?

Table 6: Random Sample of questions generated by different methods about articles from the CORD-19 corpus.

E Details of Diversity experiments

E.1 Formal definition of Diversity Metrics

For completeness, we repeat here the definitions of the diversity metrics suggested by (Shaib et al., 2024). In what follows, B is the notation used for the set of generated questions.

Definition 1 *The N-Gram Diversity (NDG) Score is defined as*

$$NDG(B) = \sum_{n=1}^{N} \frac{\#\textit{unique n-grams in } B}{\#\textit{n-grams in } B}$$

Definition 2 *The Self-Repetition Score (SR) for a natural number n, counts the fraction of questions that contain at least one n-gram that also appears in another question in the benchmark.*

Definition 3 *The Compress Ratio (CR) is the ratio between the size of the file of the benchmark, to the size of its compressed file, using* `gzip`*. Namely*

$$CR(B) = \frac{\#\textit{size of } B}{\#\textit{size of compressed } B}$$

When applied to the raw text, we refer to this metric as `word-CR`*. Conversely, we use* `PoS-CR` *to refer to the same metric applied to the Part-of-Speech tag sequence of the questions.*

Definition 4 *For a question q let v_q be the embedding vector obtained via the* `all-MiniLM-L6-v2` *sentence encoder from the Sentence Transformer package*[6]*. For questions q, q' let $sim(q, q')$ be the cosine similarity of $v_q, v_{q'}$. The Homogenization Score (HS) computes the average similarity between all question pairs in the benchmark:*

$$HS(B) = \frac{1}{|B|(|B|-1)} \sum_{q,q' \in B | q \neq q'} sim(q, q')$$

E.2 Confidence interval for diversity scores

Since the diversity metrics are not an average of point-wise scores, we had to use bootstrapping for our calculation of a confidence interval. Standard bootstrapping requires sampling with repetitions, but this would severely bias diversity metrics, especially those like NDG or SR, based on unique

[6] `https://www.sbert.net/`

```
You are given a passage of text along with a question about its content generated by an automated
process.
You must score it in multiple dimensions, as define below. For each dimension, we provide a
description of it, as well as guidelines for a numeric score. In all cases, the score is a number
between 1 and 3. Please provide your answer as a json response in the format {score name: score}
as shown in the examples below. Do not provide an explanation, just the json output.

## Fluency
Whether the question is wellformed, grammatically correct, coherent, and
fluent enough to be understood. Provide it one of 3 scores according to these guidelines

Score 1: The question is incoherent, with imprecise wording or significant grammatical errors,
making it
difficult to comprehend its meaning.
Score 2: The question is slightly incoherent or contains minor grammatical errors, but it does
not hinder the understanding of the question's meaning.
Score 3: The question is fluent and grammatically correct.


## Clarity
Whether the question is expressed clearly and unambiguously, avoiding excessive generality and
ambiguity

Score 1: The question is too broad or expressed in a confusing manner, making it difficult to
understand or leading to ambiguity. Particularly, if the generated sentence is not a question but
a declarative sentence, it should be considered in this situation.
Score 2: The question is not expressed very clearly and specifically, but it is possible to infer
the question's meaning based on the given passage.
Score 3: The question is clear and specific, without any ambiguity.


## Conciseness
Whether the question is concise and not abnormally verbose with redundant modifiers

Score 1: The question contains too much redundant information, making it difficult to understand
its intent.
Score 2: The question includes some redundant information, but it does not impact the
understanding of its meaning.
Score 3: The question is concise and does not contain any unnecessary information.


## Relevance
Whether the question is relevant to the given passage and asks for key
information from the passage

Score 1: The question is completely unrelated to the passage.
Score 2: The question is somewhat related to the passage and it asks for non-crucial information
related to the passage.
Score 3: The question is relevant to the context, and the information it seeks is crucial to the
passage.


## Consistency
Whether the information presented in the question is consistent with the passage and without any
contradictions or hallucinations

Score 1: The question contains factual contradictions with the passage or logical errors.
Score 2: The information sought in the question is not fully described in the passage.
Score 3: The information in the question is entirely consistent with the passage.
```

Figure 5: Quality assessment prompt, part 1.

n-grams. As a result, we implemented a version of bootstrapping based on sampling without repetition. We obtained 50 independent samples of the dataset, containing 80% of the data points. For each sample, we computed all the metrics, resulting in 50 scores for each method-metric pair. Then, for a given metric, we evaluated whether two methods have a statistically significant difference using t-test with $\alpha = 0.05$ over their score distributions.

```
## Answerability
Whether the question can be distinctly answered based on the passage

Score 1: The question cannot be answered based on the provided passage.
Score 2: The question can be partially answered based on the provided passage, or the answer to
the question can be inferred to some extent.
Score 3: The question can be answered definitively based on the given passage.


# Examples

Example 1
Passage: Richard ”Rick” Ducommun (July 3, 1952 - June 12, 2015) was a Canadian actor, comedian
and writer who appeared in films and television. The Burbs is a 1989 American comedy thriller
film directed by Joe Dante starring Tom Hanks, Bruce Dern, Carrie Fisher, Rick Ducommun, Corey
Feldman, Wendy Schaal and Henry Gibson. The film was written by Dana Olsen, who also has a cameo
in the movie. The film pokes fun at suburban environments and their eccentric dwellers.

Question: What star if the Burbs was Canadian?
Scores: {”fluency”: 1, ”clarity: 1, ”conciseness”: 3, ”relevance”: 3, ”consistency”: 3,
”answerability”: 1}


Example 2
Passage: At the same time the Mongols imported Central Asian Muslims to serve as administrators
in China, the Mongols also sent Han Chinese and Khitans from China to serve as administrators
over the Muslim population in Bukhara in Central Asia, using foreigners to curtail the power of
the local peoples of both lands. Han Chinese were moved to Central Asian
areas like Besh Baliq, Almaliq, and Samarqand by the Mongols where they worked as artisans and
farmers. Alans were recruited into the Mongol forces with one unit called ”Right Alan Guard”
which was combined with ”recently surrendered” soldiers, Mongols, and Chinese soldiers stationed
in the area of the former Kingdom of Qocho and in Besh Balikh the Mongols established a Chinese
military colony led by Chinese general Qi Kongzhi (Ch'i Kung-chih). After the Mongol conquest of
Central Asia by Genghis Khan, foreigners were chosen as administrators and co-management with
Chinese and Qara-Khitays (Khitans) of gardens and fields in Samarqand was put upon the Muslims as
a requirement since Muslims were not allowed to manage without them. The Mongol appointed
Governor of Samarqand was a Qara-Khitay (Khitan), held the title Taishi, familiar with Chinese
culture his name was Ahai.

Question: Where did the Mongols work?
Scores: {”fluency”: 3, ”clarity”: 2, ”conciseness”: 3, ”relevance”: 3, ”consistency”: 3,
”answerability”: 1}
```

Figure 6: Quality assessment prompt, part 2.

```
Example 3
Passage: "Domino Dancing" is a song recorded by the British synthpop duo Pet Shop Boys, released
as the lead single from their 1988 album, "Introspective". It reached number 7 on the UK Singles
Chart. Introspective is the third studio album by English synthpop duo Pet Shop Boys. It was
first released on 11 October 1988 and is the Pet Shop Boys' second-best-selling album, selling
over 4.5 million copies worldwide. (Their fifth studio album, "Very", sold more than 5 million
copies worldwide.).

Question: "Domino Dancing" is a song recorded by the British synthpop duo Pet Shop Boys, released
as the lead single from their 1988 album, "Introspective". It reached number 7 on the UK Singles
Chart, which month was the album "Introspective" first released?
Scores: {"fluency": 2, "clarity": 3, "conciseness": 2, "relevance": 3, "consistency": 3,
"answerability": 3}


Example 4
Passage: With International Criminal Court trial dates in 2013 for both President Kenyatta and
Deputy President William Ruto related to the 2007 election aftermath, US President Barack Obama
chose not to visit the country during his mid-2013 African trip. Later in the summer, Kenyatta
visited China at the invitation of President Xi Jinping after a stop in Russia and not having
visited the United States as president. In July 2015 Obama visited Kenya, as the first American
president to visit
the country while in office.

Question: Why did President Kenyatta and Deputy President William Ruto not visit the United
States in 2013?
Scores: {"fluency": 3, "clarity": 3, "conciseness": 3, "relevance": 3, "consistency": 2,
"answerability": 1}


Example 5
Passage: Even before Washington returned, Dinwiddie had sent a company of 40 men under William
Trent to that point, where in the early months of 1754 they began construction of a small
stockaded fort. Governor Duquesne sent additional french forces under Claude-Pierre Pecaudy de
Contrecœur to relieve Saint-Pierre during the same period, and Contrecœur
led 500 men south from Fort Venango on April 5, 1754. When these forces arrived at the fort on
April 16, Contrecœur generously allowed Trent's small company to withdraw. He purchased their
construction tools to continue building what became Fort Duquesne.

Question: How many men did Duquesne send to relieve Saint-Pierre?
Scores: {"fluency": 3, "clarity": 3, "conciseness": 3, "relevance": 3, "consistency": 3,
"answerability": 3}


Example 6
Passage: There are fifteen fraternities and seven sororities at the University of Chicago, as
well as one co-ed community service fraternity, Alpha Phi Omega. Four of the sororities are
members of the National Panhellenic Conference, and ten of the fraternities form the University
of Chicago Interfraternity Council. In 2002, the Associate Director of Student Activities
estimated that 8-10 percent of undergraduates were members of fraternities or sororities. The
student activities office has used similar figures, stating that one in ten undergraduates
participate in Greek life.

Question: How many fraternities are at the University of Chicago?
Scores: {"fluency": 3, "clarity": 3, "conciseness": 3, "relevance": 3, "consistency": 3,
"answerability": 3}
```

Figure 7: Quality assessment prompt, part 3.