WixQA: A Multi-Dataset Benchmark for Enterprise Retrieval-Augmented Generation

Dvir Cohen dvirco@wix.com
Lin Burg linb@wix.com
Sviatoslav Pykhnivskyi sviatoslavp@wix.com
Hagit Gur hagitg@wix.com
Stanislav Kovynov stanislavk@wix.com
Olga Atzmon olgaa@wix.com
Gilad Barkan giladba@wix.com

Wix.com AI Research, Tel Aviv, Israel

Abstract

Retrieval-Augmented Generation (RAG) is a cornerstone of modern question answering (QA) systems, enabling grounded answers based on external knowledge. Although recent progress has been driven by open-domain datasets, enterprise QA systems need datasets that mirror the concrete, domain-specific issues users raise in day-to-day support scenarios. Critically, evaluating end-to-end RAG systems requires benchmarks comprising not only question–answer pairs but also the specific knowledge base (KB) snapshot from which answers were derived. To address this need, we introduce **WixQA**, a benchmark suite featuring QA datasets precisely grounded in the released KB corpus, enabling holistic evaluation of retrieval and generation components. WixQA includes three distinct QA datasets derived from Wix.com customer support interactions and grounded in a snapshot of the public Wix Help Center KB: (i) *WixQA-ExpertWritten*, 200 real user queries with expert-authored, multi-step answers; (ii) *WixQA-Simulated*, 200 expert-validated QA pairs distilled from user dialogues; and (iii) *WixQA-Synthetic*, 6,221 LLM-generated QA pairs, with one pair systematically derived from each article in the knowledge base. We release the KB snapshot alongside the datasets under MIT license and provide comprehensive baseline results, forming a unique benchmark for evaluating enterprise RAG systems in realistic enterprise environments.

1 Introduction

Large-scale open-domain question answering (QA) datasets such as SQuAD (Rajpurkar et al., 2016), Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and InsuranceQA (Feng et al., 2015) have driven tremendous progress in AI systems based on neural networks designed for question answering. Recently, QA systems are increasingly adopting a Retrieval-Augmented Generation (RAG) framework, in which a retriever first identifies the most relevant documents and then a generator uses the retrieved text to produce candidate answers (Lewis et al., 2021). However, many real-world settings—especially in enterprise customer support—demand domain-specific evaluation, robust retrieval from curated knowledge bases, and step-by-step guidance rather than a one-shot answer. Central to these real-world scenarios is the need to synthesize information from multiple documents to fully address user queries, a capability that our datasets are designed to evaluate.

Enterprise queries present unique challenges, often demanding complex procedural guidance and specialized vocabulary (characteristics of Long-form Question Answering, LFQA), rather than the short, factual answers targeted by many existing QA benchmarks (Rajpurkar et al., 2016; Kwiatkowski et al., 2019; Joshi et al., 2017). To address this specific need, we introduce **WixQA**, a benchmark suite designed for the enterprise domain. Recognizing the limitations of existing resources, WixQA adopts a two-pronged design. This approach explicitly incorporates datasets featuring long-form, multi-step answers crucial for troubleshooting and task resolution in enterprise contexts, alongside datasets with shorter, specific answers reflecting more straightforward real-world scenarios. Such diversity makes our benchmark especially versatile for developing robust QA models adept at handling a wide spectrum of response types.

WixQA realizes this vision with three complementary datasets derived from Wix.com support interactions:

- **WixQA-ExpertWritten:** A collection of 200 genuine customer queries paired with step-by-step answers manually authored and validated by domain experts, reflecting real-world support challenges.

- **WixQA-Simulated:** A set of 200 examples distilled from multi-turn user–chatbot dialogues into clear, single-turn QA pairs, each meticulously validated for procedural accuracy by domain experts through simulation.
- **WixQA-Synthetic:** Comprising 6,221 question–answer pairs automatically extracted using Large Language Models (LLMs) from Wix articles, this dataset offers scale and diversity for training robust retrieval models.

Recent works such as TechQA (Castelli et al., 2019), FinTextQA (Chen et al., 2024), FINQA (Chen et al., 2022), PubMedQA (Jin et al., 2019), WikiHowQA (Bolotova-Baranova et al., 2023), and AmazonQA (Gupta et al., 2019) underscore the trend toward benchmarks that capture expert curation, automatic extraction, and domain-specific nuances. Their diverse scales and focuses further motivate our enterprise-scale approach.

Our key contributions are as follows:

- **A Diverse Suite of Enterprise QA Datasets:** Introduction of **WixQA**, comprising three complementary, KB-grounded datasets (ExpertWritten, Simulated, Synthetic) reflecting realistic enterprise support interactions and varying curation methods.
- **Multi-article dependency:** A unique aspect of our datasets is that both ExpertWritten and Simulated datasets' answers can be based on more than one article within the Wix knowledge base. This adds complexity to the RAG process, as the model must retrieve and synthesize information from multiple sources to generate a comprehensive and accurate response. This multi-article dependency reflects real-world scenarios where user queries require integration of knowledge from various parts of the corpus.
- **Enterprise-Scale Knowledge Base:** We release a unified corpus of 6,221 Wix help articles, forming a domain-specific knowledge base that reflects the nuanced language and multi-step procedures typical of enterprise support workflows.
- **Comprehensive Benchmarking:** Extensive experiments across all three datasets provide in-depth analyses of retrieval and generation performance in realistic enterprise environments, highlighting the challenges and opportunities for further advancements.

In the following sections, we review related work (§2), detail our data collection and annotation protocols in addition to comprehensive statistics for each dataset (§3), and showcase baseline experiments (§4). We conclude with a discussion of findings and directions for future work (§5).

2 Related Work

Recent advancements in question answering (QA) have seen a significant shift from general-purpose, Wikipedia-based benchmarks toward specialized, domain-specific datasets. This shift is motivated by the increasing demand for practical applications in enterprise environments, where retrieval accuracy, domain-specific vocabulary, and procedural or multi-step reasoning play critical roles.While earlier benchmarks (e.g., SQuAD (Rajpurkar et al., 2016), TriviaQA (Joshi et al., 2017), Natural Questions (Kwiatkowski et al., 2019)) centered on open-domain factual knowledge from Wikipedia, real-world QA increasingly requires domain-adapted datasets.

Several notable QA benchmarks have emerged addressing needs in organizational and enterprise contexts, focusing on customer support, internal documentation, and procedural guidance. TechQA (Castelli et al., 2019) is particularly relevant, comprising real user-generated questions from IBM's technical support forums paired with answers located within IBM's corpus of technical documents (Technotes). It provides both short and long-form answers, making it suitable for evaluating retrieval accuracy within an enterprise technology support context. Similarly, Doc2Dial (Feng et al., 2020) and MultiDoc2Dial (Feng et al., 2021) focus on customer support dialogues grounded in domain-specific documents (sourced from public sector entities such as government agencies), evaluating conversational RAG systems that track information over multiple turns. These benchmarks highlight the importance of using authentic *organizational data and domain relevance*, a principle shared by our WixQA datasets, which are exclusively derived from Wix customer support interactions and its carefully curated repository of Wix articles. This ensures evaluation on realistic enterprise support scenarios involving multi-step procedures and domain-specific terminology.

Procedural knowledge is another key focus area. WikiHowQA (Bolotova-Baranova et al., 2023) leverages WikiHow articles to evaluate QA sys-

Dataset	Domain	Questions Source	Answers	Knowledge Base	Size
TechQA	Tech Support	Expert-selection	Expert-written	IBM Technotes	600
BioASQ-QA	Biomedical	Manual	Expert-written	PubMed	500
FinTextQA	Finance	Auto-generated	Auto-generated	Finance Textbooks	1,262
FinQA	Finance	Expert-curated	Expert-written	Financial Docs	8,200
EmrQA	Clinical	Auto-extracted	Auto-extracted	EMRs	2M
DomainRAG	Education	Hybrid	Hybrid	University Pages	395
Doc2Dial	Gov. Services	Dialogues	Expert-written	Domain Docs	4,500
WikiHowQA	Procedural	Expert-selected	Expert-written	WikiHow Articles	11,746
DoQA	Community FAQs	Human-curated	Human-curated	Stack Exchange	10,917
ConvRAG	Conversational	Hybrid	Expert-validated	Conversations	1,000
WixQA-EW	Enterprise	User queries	Expert-written	Wix Articles	200
WixQA-Simulated	Enterprise	User queries	Expert-validated	Wix Articles	200
WixQA-Synthetic	Enterprise	Auto-generated	Auto-generated	Wix Articles	6,221

Table 1: Extended Overview of QA Benchmarks including WixQA; for datasets with train–test splits, the reported size refers to the test set

tems on step-by-step instructions. Although not strictly organizational data, its procedural nature aligns closely with enterprise troubleshooting scenarios, emphasizing detailed, actionable answers. Datasets like DoQA (Campos et al., 2020) have introduced FAQ-style conversational QA across multiple Stack Exchange domains, reflecting realistic interactions in specialized communities and the complexity of retrieving multi-step answers.

The quality and curation method of answers vary across benchmarks. Expert curation is a hallmark of quality in specialized QA benchmarks like TechQA and FinQA (Chen et al., 2022), which rely heavily on expert-validated answers. Doc2Dial and MultiDoc2Dial provide annotated dialogues grounded in organizational documents. Synthetic datasets, such as FinTextQA (Chen et al., 2024) (focused on finance), leverage automatic methods for large-scale coverage. WixQA adopts a multi-faceted approach: WixQA-ExpertWritten and WixQA-Simulated feature rigorous expert validation to ensure procedural accuracy, similar to TechQA and Doc2Dial. Meanwhile, WixQA-Synthetic, auto-generated from expert-written Wix articles, offers extensive training data while maintaining high preliminary accuracy, balancing scale with precision akin to synthetic benchmarks. Further expanding the evaluation landscape, datasets like DomainRAG (Wang et al., 2024b) and ConvRAG (Ye et al., 2024) incorporate hybrid curation methods and address aspects like multilingual and conversational RAG, highlighting the growing diversity in benchmark design

The supporting knowledge base characteristics significantly impact RAG performance. Domain-focused benchmarks such as Doc2Dial, Multi-Doc2Dial, and FinQA highlight how curated organizational documents (like customer support documents or financial statements) enhance retrieval precision and answer relevance. WixQA similarly provides a unified corpus of 6,221 Wix articles, forming a cohesive, domain-specific knowledge base. This approach rigorously evaluates models' ability to retrieve highly relevant procedural content from organizational documentation, a crucial capability for enterprise RAG systems.

The WixQA datasets introduced in this work uniquely contribute to this landscape by blending expert-written (ExpertWritten), expert-validated distilled (Simulated), and LLM-generated (Synthetic) QA pairs, all derived from Wix's internal knowledge base. A key distinction is the explicit inclusion of multi-article dependencies in the ExpertWritten and Simulated datasets, requiring synthesis across documents—a common real-world enterprise challenge. Table 1 overviews these benchmarks alongside WixQA, highlighting their diverse characteristics.

3 Data Collection and Annotation

This section describes the methodology for creating our three datasets and our knowledge base. Each dataset was designed to address different aspects of enterprise QA. The creation of answers, particularly those involving expert curation, adhered to specific instructions (see Appendix A) aimed at ensuring responses were comprehensive, KB-grounded, and user-centric.

3.1 Knowledge Base

The Knowledge Base (KB) supporting our benchmark contains a snapshot of English-only articles from Wix's customer support repository[1]. The KB consists of 6,221 articles distributed across three distinct types:

- **Article (66%):** General-purpose articles covering a wide range of topics, including tutorials, troubleshooting guides, and feature explanations.
- **Feature Request (32%):** Articles that provide information about unsupported features, allow users to vote for them, and announce when a feature has been implemented.
- **Known Issue (1%):** Articles that document known problems acknowledged by Wix and provide updates on their resolution status.

The Knowledge Base (KB) is a crucial component of RAG applications, providing the necessary context to ground the LLM's responses. This ensures answers rely on specific, up-to-date enterprise information, particularly where the LLM's internal knowledge may be absent, outdated, or limited.

Table 2 summarizes key statistics for the three WixQA datasets, highlighting differences in scale, question/answer complexity, and multi-article dependency.

The higher multi-article percentage in ExpertWritten (27%) compared to Simulated (14%) reflects the former's design for comprehensive expert solutions versus the latter's focus on concise, distilled answers.

3.2 WixQA-ExpertWritten - Real User Queries with Expert Answers

WixQA-ExpertWritten, the first dataset in our suite, contains 200 authentic customer queries from Wix.com support interactions. Each query is paired with a detailed, step-by-step answer meticulously authored by domain experts. An example of a question and an answer is provided in Appendix B.1. Key characteristics and the creation process include:

- **Source of Questions:** Genuine user queries submitted through Wix support channels, covering a diverse range of real-world issues such as domain configuration, SSL certificate troubleshooting, editor functionalities, etc.
- **Expert Answer Curation:** Ground truth answers were manually authored by Wix support experts. These answers are designed to be comprehensive and granular, often providing detailed step-by-step instructions necessary for resolving complex user problems and ensuring users have a clear, actionable solution. This authoring process followed the specific instructions outlined in Appendix A, which emphasize fidelity to KB content, completeness, and user-focused communication.
- **Knowledge Base Grounding and Multi-Article Synthesis:** Answers are grounded in the official Wix knowledge base (§3.1). Crucially, many answers require synthesizing information from multiple KB articles (27% involve more than one article) to fully address the user's query, reflecting realistic enterprise support scenarios where information is often distributed across documents.
- **Rigorous Manual Validation:** A multi-stage validation process ensured answer quality and accuracy. Initially drafted answers were reviewed by three peer experts; acceptance required a majority vote. Subsequently, two senior experts conducted a final validation across the entire dataset. Answers failing to meet strict criteria for factual accuracy, clarity, and relevance to the query were removed at either stage. This rigorous protocol ensures the dataset reflects production-level support standards.
- **Intended Use Case:** This dataset is ideal for evaluating the ability of RAG systems to handle authentic, potentially complex user queries that necessitate detailed, multi-step, and sometimes multi-source answers. It tests the system's capacity for both accurate retrieval (potentially across multiple documents) and comprehensive generation.

[1] https://support.wix.com/en

Dataset	Size	Question Tokens	Answer Tokens	Multi-Article %
ExpertWritten	200	19	172	27%
Simulated	200	12	50	14%
Synthetic	6,221	24	130	0% (by design)

Table 2: Key Statistics for WixQA Datasets (Median Values)

- **Statistics:** As summarized in Table 2, this dataset features concise real user queries (median 19 tokens) paired with comprehensive, expert-authored answers (median 172 tokens). This difference highlights the need for detailed, step-by-step guidance. The significant context size (median 5,928 tokens per answer's source articles) and multi-article ratio (27% require >1 article) underscore the complexity addressed.

3.3 WixQA-Simulated: Expert-Validated QA Pairs from User Conversations

Complementing the *WixQA-ExpertWritten* dataset, *WixQA-Simulated* offers 200 QA pairs between Wix users and support chatbot dialogues, refined for conciseness and procedural accuracy via expert validation. It targets evaluating RAG systems on generating comprehensive yet effective guidance. An example is provided in Appendix B.2.

The creation process involved several key stages:

- **Source Material:** We collected multi-turn conversational dialogues between Wix users and the support chatbot addressing specific issues.
- **QA Pair Distillation:** Using an LLM, these dialogues were distilled into single-turn question–answer pairs. The objective was to capture the core user problem and the essential steps of the expert's solution concisely.
- **Knowledge Base Grounding:** Answers were grounded in the Wix knowledge base (described in §3.1). Similar to *WixQA-ExpertWritten*, answers may require synthesizing information from multiple KB articles, but the emphasis here is on brevity and directness in the final distilled answer.
- **Rigorous Multi-Stage Validation:** The distilled QA pairs underwent a stringent validation process. First, automatic filtering removed irrelevant QA (e.g., requests for human agents, ambiguous queries, questions unanswerable solely by the KB). Second, three domain experts manually reviewed the remaining pairs, discarding those where the answer did not appropriately address the question. Finally, for the pairs passing these filters, annotators performed simulation-based validation: they meticulously followed the step-by-step instructions in each answer to verify its correctness and confirm that it resolved the user's specific problem. This multi-stage process ensured the final set of 200 QA pairs is both highly relevant and procedurally accurate.
- **Intended Use Case:** This dataset serves as a benchmark for evaluating a RAG system's ability to generate specific, concise, and accurate responses, particularly for procedural or multi-step tasks. It contrasts with *WixQA-ExpertWritten*'s focus on comprehensive answers, instead testing for maximal accuracy within minimal length.
- **Statistics:** Designed for conciseness, this dataset features short questions (median 12 tokens) and relatively brief answers (median 50 tokens), as shown in Table 2. This contrasts with the ExpertWritten dataset and reflects the goal of evaluating accurate, distilled guidance. A notable portion (14%) still requires multi-article synthesis.

3.4 WixQA-Synthetic: Large-Scale QA Pairs via LLM-Based Extraction

To complement the expert-curated datasets and provide large-scale data suitable for training robust models, we created *WixQA-Synthetic*. This dataset comprises 6,221 question–answer pairs, with one pair generated for each of the 6,221 articles in the Wix.com knowledge base.An example is provided in Appendix B.3.

The generation and validation process involved the following steps:

- **Automated QA Generation:** We applied a state-of-the-art LLM, specifically GPT-4o, to each of the 6,221 articles within our knowledge base (§3.1). Recognizing the distinct structure and purpose of the three article types (Article, Feature Request, Known Issue), we employed tailored prompts for each type to optimize the quality of the extracted QA pairs. An example prompt for the 'Feature Request' type is provided in Appendix E.

- **Single-Article Grounding:** By design, each generated question–answer pair is explicitly linked to the single source knowledge base article from which it was derived. This provides a direct ground truth reference (article URL) for evaluating retrieval performance using common metrics like Precision@K and Recall@K. This single-source grounding contrasts with the potential multi-article synthesis required for the *WixQA-ExpertWritten* and *WixQA-Simulated* datasets.
- **Quality Assurance and Validation:** To assess the quality of the LLM-generated data, two domain experts manually evaluated a random sample of 250 QA pairs. This review confirmed high fidelity, with over 90% of the sampled answers found to be correct and relevant to the source article. Building on this positive result, we performed manual sanity checks across the entire generated set (6,221 pairs) to ensure overall data integrity and absence of major formatting errors or inconsistencies. This multi-faceted validation confirmed the reliability of the extraction pipeline.
- **Intended Use Case:** With its substantial size, this dataset is particularly well-suited for training or fine-tuning the retrieval and generation components of RAG systems. It offers broad coverage of the knowledge base content, enabling models to learn domain-specific patterns and terminology at scale.
- **Statistics:** With 6,221 QA pairs, this dataset offers considerable volume for training (Table 2). Median token counts are 24 for questions and 130 for answers, aligning with the complexity of the curated WixQA datasets while providing greater quantity.

3.5 Data Availability

The WixQA datasets and KB corpus are publicly available as a Hugging Face Datasets [2] to encourage enterprise RAG research.

4 Experiments

In this section, we establish a benchmark baseline for retrieval-augmented generation (RAG) in enterprise support settings. Our evaluation spans the three distinct datasets (ExpertWritten, Simulated, and Synthetic) and leverages a combination of classical and dense retrieval methods alongside multiple state-of-the-art generation models. The goal of this benchmark creation is to provide a robust foundation for future research in procedural, multi-document QA. The pipeline was implemented using the FlashRAG tool (Jin et al., 2024), and executed in a single run, with specific configuration parameters detailed in Appendix F.

4.1 Retrieval Methods

We employed two standard retrieval approaches: keyword-based BM25 (Robertson and Zaragoza, 2009) and semantic-based E5 dense retrieval (specifically, `e5-large-v2`, 335M parameters) (Wang et al., 2024a). For retrieval, the top $k = 5$ documents were selected.

4.2 Generation Models

Our benchmark tests generation components that synthesize retrieved context into coherent, detailed procedural answers. We evaluate several state-of-the-art models—Claude 3.7, Gemini 2.0 Flash, GPT-4o, and GPT-4o Mini—running each with its default configuration to ensure a fair comparison across diverse architectures.

These models vary in language modeling and reasoning, affecting fluency, coherence, and multi-step explanation ability. This comparison reveals strengths and trade-offs for generating procedural responses aligned with real-world queries.

4.3 Evaluation Metrics

Due to the multi-faceted and procedural nature of the answers, we rely on a diverse set of evaluation metrics:

- **F1:** Token-level F1 score computed between the generated and gold answers.
- **BLEU:** An n-gram overlap metric that quantifies the similarity between the generated and the gold answers (Papineni et al., 2002).
- **ROUGE-1 and ROUGE-2:** Metrics that capture unigram and bigram overlaps, respectively, to assess answer adequacy (Lin, 2004).
- **Factuality:** An LLM-based judge metric (Tan et al., 2025) evaluating the factual alignment

[2] https://huggingface.co/datasets/Wix/WixQA

Parameters		Metrics					
Retrieval Model	Generation Model	F1	BLEU	ROUGE-1	ROUGE-2	Context Recall	Factuality
bm25	claude 3.7	0.37	0.09	0.31	0.12	0.73	0.80
bm25	gemini 2.0 flash	0.39	0.12	0.32	0.15	0.72	0.72
bm25	gpt 4o	0.36	0.08	0.29	0.12	0.73	0.83
bm25	gpt 4o mini	0.37	0.08	0.30	0.11	0.72	0.76
e5	claude 3.7	0.39	0.10	0.32	0.13	**0.81**	0.82
e5	gemini 2.0 flash	**0.43**	**0.14**	**0.35**	**0.17**	**0.81**	0.76
e5	gpt 4o	0.37	0.08	0.30	0.12	**0.81**	**0.85**
e5	gpt 4o mini	0.39	0.09	0.31	0.12	**0.81**	0.79

Table 3: Performance on the **ExpertWritten** dataset

Parameters		Metrics					
Retrieval Model	Generation Model	F1	BLEU	ROUGE-1	ROUGE-2	Context Recall	Factuality
bm25	claude 3.7	0.21	0.03	0.21	0.08	0.55	0.74
bm25	gemini 2.0 flash	0.28	0.04	0.26	0.11	0.55	0.63
bm25	gpt 4o	0.22	0.03	0.22	0.08	0.55	0.76
bm25	gpt 4o mini	0.23	0.03	0.21	0.08	0.55	0.74
e5	claude 3.7	0.23	0.04	0.21	0.08	**0.67**	**0.77**
e5	gemini 2.0 flash	**0.30**	**0.05**	**0.28**	**0.12**	**0.67**	0.66
e5	gpt 4o	0.24	0.03	0.23	0.09	**0.67**	**0.77**
e5	gpt 4o mini	0.24	0.04	0.23	0.09	**0.67**	0.75

Table 4: Performance on the **Simulated** dataset

Parameters		Metrics					
Retrieval Model	Generation Model	F1	BLEU	ROUGE-1	ROUGE-2	Context Recall	Factuality
bm25	claude 3.7	0.42	0.18	0.38	0.22	0.95	0.84
bm25	gemini 2.0 flash	**0.59**	**0.33**	**0.53**	**0.37**	0.96	0.85
bm25	gpt 4o	0.47	0.20	0.42	0.24	0.95	0.86
bm25	gpt 4o mini	0.43	0.17	0.38	0.19	0.95	0.84
e5	claude 3.7	0.43	0.18	0.38	0.22	**0.97**	0.83
e5	gemini 2.0 flash	0.58	**0.33**	**0.53**	0.36	**0.97**	0.84
e5	gpt 4o	0.48	0.21	0.44	0.26	0.96	**0.87**
e5	gpt 4o mini	0.44	0.18	0.39	0.20	**0.97**	0.86

Table 5: Performance on the **Synthetic** dataset

between the generated answer and the ground truth answer (see Appendix C). The LLM judge receives the original query, the generated answer, and the ground truth answer, and produces a score on a 0-1 scale based on how accurately the generated answer reflects the essential factual information present in the ground truth. This assesses whether the generator accurately utilized the provided context and avoided introducing factual errors or hallucinations. For our experiments, we utilized GPT-4o as the LLM judge for this metric.

- **Context Recall:** An LLM-based judge metric assessing the relationship between the retrieved context and the ground truth answer. To evaluate this, an LLM judge is provided with the user query, the retrieved context, and the ground truth answer. Following a specific instructional prompt (see Appendix D), the LLM's task is to evaluate the extent to which the essential information required to formulate the ground truth answer is present within the retrieved context. It breaks down the ground truth answer into its core informational components and checks for their presence in the context, ignoring any additional, non-essential information within the context itself. Based on this analysis of information coverage, the LLM assigns a score on a 0-1 scale, reflecting the degree to which the retrieved context contains the necessary information to construct the ground truth answer. For our experiments, we utilized GPT-4o as the LLM judge for this metric as well.

4.4 Results and Benchmark Baseline

Tables 3, 4, and 5 summarize the performance of the RAG pipeline on the ExpertWritten, Simulated, and Synthetic datasets, respectively. Our baseline results yield several key insights:

- **Dense Retrieval Boosts Recall for Complex Queries:** The E5 dense retriever consistently outperforms BM25 on Context Recall, particularly for the ExpertWritten and Simulated datasets requiring multi-article synthesis. This highlights the benefit of semantic matching for complex information needs.
- **Dataset Difficulty Varies:** Performance differs markedly across datasets, with Synthetic yielding the highest scores, followed by ExpertWritten, and then Simulated proving the most challenging. This suggests varying difficulty levels related to query authenticity, answer complexity, and grounding requirements (single vs. multi-article).
- **Generation Models Exhibit Trade-offs:** No single generator excels universally. Models show distinct performance profiles, with some favouring n-gram similarity (F1, BLEU, ROUGE) while others achieve higher Factuality scores, indicating clear trade-offs relevant to specific application goals.
- **Enterprise RAG Requires Further Advancement:** While viable, the baseline scores, especially on ExpertWritten and Simulated datasets, underscore the significant challenge of generating accurate procedural answers from enterprise knowledge and highlight the need for continued research using benchmarks like WixQA.

These baselines show our RAG pipeline's viability for complex enterprise queries and offer a strong benchmark for future retrieval and generation research.

5 Conclusion and Future Work

To advance enterprise Retrieval-Augmented Generation (RAG) systems, we introduced **WixQA**, a benchmark suite featuring three QA datasets (ExpertWritten, Simulated, and Synthetic) and a 6,221-article knowledge base. WixQA's enables assessment of domain-specific procedural QA, particularly tasks requiring multi-document synthesis and complex, long-form answers. Our comprehensive baseline experiments establish initial performance levels and, critically, highlight persistent challenges in enterprise RAG.

Future work will focus on scaling these datasets, introducing multi-hop retrieval tasks, and refining human evaluation protocols. We release WixQA and the associated knowledge base (publicly available on Hugging Face) to foster advancements in RAG systems for reliable, user-centric enterprise applications and to provide a strong baseline for future research.

Acknowledgments

We would like to extend our sincere gratitude to the dedicated efforts of our dataset labelers, including the Wix.com Customer Care Experts. Their meticulous attention to detail and unwavering commitment have been instrumental in refining the quality and accuracy of our datasets. We deeply appreciate the contributions of Serhii Shytikov, Kseniia Nalyvaiko, Kateryna Kharlamova, Anastasiia Bikovska, Inha Zatoka, Roy Shalish, and Ariel Yaakobi, whose expertise and hard work have significantly enhanced this research.

References

Valeriia Bolotova-Baranova, Vladislav Blinov, Sofya Filippova, Falk Scholer, and Mark Sanderson. 2023. WikiHowQA: A comprehensive benchmark for multi-document non-factoid question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5291–5314, Toronto, Canada. Association for Computational Linguistics.

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. Doqa – accessing domain-specific faqs via conversational qa. *Preprint*, arXiv:2005.01328.

Vittorio Castelli, Rishav Chakravarti, Saswati Dana, Anthony Ferritto, Radu Florian, Martin Franz, Dinesh Garg, Dinesh Khandelwal, Scott McCarley, Mike McCawley, Mohamed Nasr, Lin Pan, Cezar Pendus, John Pitrelli, Saurabh Pujar, Salim Roukos, Andrzej Sakrajda, Avirup Sil, Rosario Uceda-Sosa, and 2 others. 2019. The techqa dataset. *Preprint*, arXiv:1911.02984.

Jian Chen, Peilin Zhou, Yining Hua, Loh Xin, Kehui Chen, Ziyuan Li, Bing Zhu, and Junwei Liang. 2024. Fintextqa: A dataset for long-form financial question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 6025–6047. Association for Computational Linguistics.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2022. Finqa: A dataset of numerical reasoning over financial data. *Preprint*, arXiv:2109.00122.

Minwei Feng, Bing Xiang, Michael R. Glass, Lidan Wang, and Bowen Zhou. 2015. Applying deep learning to answer selection: A study and an open task. *Preprint*, arXiv:1508.01585.

Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. MultiDoc2Dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6162–6176, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128, Online. Association for Computational Linguistics.

Mansi Gupta, Nitish Kulkarni, Raghuveer Chanda, Anirudha Rayasam, and Zachary C Lipton. 2019. Amazonqa: A review-based question answering task. *Preprint*, arXiv:1908.04364.

Jiajie Jin, Yutao Zhu, Xinyu Yang, Chenghao Zhang, and Zhicheng Dou. 2024. Flashrag: A modular toolkit for efficient retrieval-augmented generation research. *CoRR*, abs/2405.13576.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *Preprint*, arXiv:1909.06146.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *Preprint*, arXiv:1705.03551.

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, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *Preprint*, arXiv:1606.05250.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3:333–389.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2025. Judgebench: A benchmark for evaluating llm-based judges. *Preprint*, arXiv:2410.12784.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024a. Text embeddings by weakly-supervised contrastive pre-training. *Preprint*, arXiv:2212.03533.

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

Linhao Ye, Zhikai Lei, Jianghao Yin, Qin Chen, Jie Zhou, and Liang He. 2024. Boosting conversational question answering with fine-grained retrieval-augmentation and self-check. *Preprint*, arXiv:2403.18243.

A Instructions for Expert Answer Authoring

When authoring answers based on Knowledge Base (KB) articles, you were to adhere to the following instructions:

- **Provide Relevant KB Article URLs:** Identify and provide the URLs of all relevant KB articles that, in combination, fully answer the user's question. For data collection, each URL was to be listed in a separate column.
- **Write a Full and Relevant Answer:** Compose a full answer. A full answer provides only the relevant information, and all the relevant information, necessary for users to resolve their issues. Information not pertinent to the solution should be omitted.
- **Adhere Strictly to KB Content:** Copy the answer from the identified KB articles as much as possible.
- **Exclusive Reliance on KB:** Do not use any information from your previous experience that does not rely on the provided KB.
- **User-Centric Phrasing and Linking:** Phrase your answer as you would to a real user. When needed and supported by the KB material, add links within the answer (e.g., to other KB articles, specific Wix resources, or pertinent external sources).

B QA Examples

B.1 WixQA-ExpertWritten Example

Below is a representative example of a QA in the WixQA-ExpertWritten dataset:

QUESTION

I cant access my old email address to change my password. I have a recovery phone number added to my Wix account.

ANSWER

If you've previously added a Recovery Phone Number to your Wix account, you can get a one-time SMS code sent to your mobile phone to recover your Wix account.

To recover your account:

1. Go to the Wix Log In page. (https://users.wix.com/signin)

2. Click Forgot Email?.

3. Select Can't access my login email.

4. Click Next.

5. Enter the email address associated with your Wix account.

6. Click Next.

7. Select Send me a code to my phone number ending with XXXX and click Next.

8. Enter the one-time verification code sent to your phone under Enter code here.

9. Click Verify.

10. Enter a new password and retype it.

11. Click Save New Password.

After you recover your account, you should change the email address of your account (https://support.wix.com/en/article/changing-the-email-address-of-your-wix-account) to one that you are able to access.

To change your email address:

1. Go to Account Settings in your Wix account.

2. Click the Account email field.

3. Enter your Wix account password in the pop-up.

4. Enter your new account email address.

5. Retype your new email account email address to confirm it.

6. Click Next.

7. Check your email account for an email from Wix.

8. Copy the 6-digit confirmation code.

9. Go back to Account Settings in your Wix account.

10. Paste the code into the pop-up.

11. Click Submit.

ARTICLES

"391586...", "e93664..."

B.2 WixQA-Simulated Example

Below is a representative example of a QA in the WixQA-Simulated dataset:

QUESTION

How do I link a button to the users member page without linking to a dataset?

ANSWER

To link a button to a members page without using a dataset, add a button in the Wix Editor, click the Link icon, select Page, choose the members page from the list, and click Done.

ARTICLES

"926ba4...", "fc77f8..."

B.3 WixQA-Synthetic Example

Below is a representative example of a QA in the WixQA-Synthetic: dataset:

QUESTION

Why is my blog post showing the full content instead of just a preview when connected to pricing plans?

ANSWER

> This is a known problem where blog posts connected to pricing plans show the full post instead of the preview. As a temporary workaround, you can remove the preview entirely by following these steps:
>
> 1. Enter Post composer for Post
>
> 2. Click monetize
>
> 3. Move preview to the top to remove the entire preview
>
> 4. Republish post
>
> 5. Refresh live site
>
> ARTICLES
>
> "3835e0..."

C LLM Prompt for Factuality Evaluation

Below is the prompt used to evaluate the factual consistency (i.e., actuality) between the generated answer and the golden answer:

> ROLE
>
> You are a Factual Alignment Expert. Your job is to evaluate how well an AI response includes the essential information from a ground truth answer (GT answer) according to a given user query.
>
> Note that the Ground Truth (GT Answer), is the "Correct" answer generated by an expert, and was created to evaluate the model, and is NOT part of the AI response or the context.
>
> TASK DESCRIPTION
>
> You will be presented with three elements: a question, a GT answer, and an AI response. Determine how well the AI response includes the essential information from the GT answer that helps to solve the user's query. In case of any additional or extra information present in the AI response, only make sure it's not preventing the user from solving his query.
>
> EVALUATION CRITERIA
>
> 5: Complete Match - All essential information from GT answer appears in AI response, providing complete solution to the query
>
> 4: Strong Match - Most essential information is present, with only minor details missing that don't impact the solution significantly
>
> 3: Partial Match - Core information is present but missing some important details that would help better solve the query
>
> 2: Limited Match - Only basic or partial information present, missing several essential elements needed for the solution
>
> 1: Poor Match - Missing most essential information or contains incorrect information that could mislead the user
>
> INSTRUCTIONS
>
> 1. Read the question carefully and analyze the ground truth answer to identify all key information elements that help solve the query
>
> 2. Compare the AI response (candidate answer) against the ground truth, focusing on presence of important information
>
> 3. Evaluate the completeness and accuracy of the information transfer
>
> 4. Assign a rating (0-1) based on how well important information is preserved
>
> 5. Provide a brief explanation focusing on factuality

D LLM Prompt for Context Recall Evaluation

Below is the prompt used to evaluate the context recall between the retrieved context and the golden answer:

> ROLE You are a Context Evaluation Expert. Your job is to assess how well a retrieved context contains the essential information present in a ground truth answer (GT answer).
>
> TASK DESCRIPTION You will be presented with three pieces of information: a user query, its ground truth answer , and a retrieved context (that will be used to create an AI response from). Determine how well the essential information from the GT answer appears in the retrieved context. Additional information in the retrieved context should not affect the scoring.
>
> EVALUATION CRITERIA
>
> 5: Complete Match - All essential information from the GT answer is present in the retrieved context. The context fully enables answering the user's question. 4: Strong Match - All essential information is present, but some minor details are missing. The context still effectively answers the user's question. 3: Partial Match - Most essential information is present, but some important details are missing. The context partially answers the user's question. 2: Weak Match - Only basic or limited essential information is present. The context provides insufficient information to properly answer the user's question. 1: No Match - Essential information is missing or incorrect. The context cannot be used to answer the user's question.
>
> INSTRUCTIONS
>
> 1. Read the question to understand the idea of what the user asks for 2. Break down the GT answer into essential information (key facts, main concepts, direct answers). 3. Check if these information pieces appear in the retrieved context 4. Focus only on finding the ground truth information in the context - ignore any additional or extra information present in the retrieved context 5. Assign a rating (0-1) based on information coverage and relevance

E LLM Prompt for QA pairs extraction from Feature Request articles

Below is a prompt for GPT-4o we used to extract question–answer pairs for Feature Request type of articles:

> The kb article provided below contains information about a feature or an action that may not be

implemented yet. Find the not implemented feature from the article, ask a plausible user question if the feature is supported, and answer it using the article.

Ensuring the following:

1. Use the exact wording from the article for the answer whenever possible.

2. Avoid any hallucinations at the end, such as "For more information..."

3. If there are relevant step-by-step instructions, include them in the answer.

4. Do not skip the information from the step-by-step instructions.

5. If there is a relevant workaround or tip, include it.

6. Copy relevant links directly from the context as they are.

7. Do not include phrases suggesting to contact support unless absolutely necessary.

8. Ensure that the image names are not used as usual text.

9. Do not include: "We are always working to update and improve our products, and your feedback is hugely appreciated".

10. Do not include announcement like "We are excited to announce".

Provide the output as JSON with "question" and "answer" fields. Format the "answer" field value as a markdown.

F FlashRAG Configuration Parameters

This section lists the core FlashRAG configuration parameters used, corresponding to the framework's default settings. Parameters related to file paths and environment specifics have been omitted. Note that specific hyperparameters tuned during our experiments (such as the generation model) are not detailed here.

- generator_batch_size: 25
- generator_max_input_len: 50,000
- retrieval_topk: 5
- retrieval_query_max_length: 10,000
- retrieval_batch_size: 1024
- max_tokens: 1024
- temperature: 0