Continuous Diffusion-based Text Generation for Query Expansion in LM, an Intuition

Avi Bleiweiss
BShalem Research
Sunnyvale, CA, USA
avibleiweiss@bshalem.onmicrosoft.com

Abstract

Large Language Models (LLMs) have been used to generate textual expansion of original queries for improving information retrieval (IR) and question answering (QA) responses. The recent emergence of retrieval augmented generation had LLM-based query expansions (QE) more faithful and grounded to document corpus. Our query expansion methodology employs a two-step Chain-of-Thought (CoT) prompting in an instruction-tuned LLM. In this framework, we propose a rendition of a continuous diffusion-based generative QE model in LM. Rather than replicating a single query instance, we utilize a forward diffusion process to dynamically perturb the initial query with fine-grained Gaussian noise, aiming to improve QE theoretical acceptance. In practice, we post-train an LSTM neural network and generate a query every epoch. We evaluated our method contrasting table QA with text QA tasks on two open-domain question answering datasets, respectively: a) FeTaQA, a relatively new dataset with question-answer pairs assembled from high quality descriptions of Wikipedia tables, and b) AmazonQA, a large-scale review-based dataset that spans a broad gamut of product categories. On FeTaQA, a free-form generative table QA, we assessed the agreement of two relevance raters using Kappa scoring, and on AmazonQA, we studied relational reasoning across product domains using native embedding representation for computing similarity.

1 Introduction

Query expansion (QE) has been a widely adopted technique in information search applications (Rocchio 1971) that augments the original query with additional contexts to match target documents. Earlier studies used originally retrieved documents as pseudo-relevance feedback (PRF; Robertson 1990), however, the effectiveness of these methods were limited by the quality of initial retrievals. In contrast, document expansion applies similar techniques but expands document terms throughout indexing rather than query terms during retrieval (Nogueira et al. 2019).

Prior research (Claveau 2022) explored the use of neural text generation to expand queries and proved the provided new terms to the query are also a better estimate of their relative weights. The main application domains to apply query augmenting in LLMs include information retrieval (IR) and question answering (QA). Recently, LLMs has been a prominent area of interest and have been utilized to generate query expansions with their intrinsic knowledge. Prompt tuning to control the model textual response, further concatenated with the initial query has been a widespread and simple approach adopted for query expansion. Wang, Yang, and Wei (2023); Jagerman et al. (2023) studied a variety of different prompts, including zero-shot, few-shot and Chain-of-Thought (CoT; Wei et al. 2022b). Evidently, CoT prompts are especially effective for query expansion as they generate a series of intermediate reasoning steps and provide a large number of terms related to the initial query.

Llama 3.1 (Meta 2024) is a foundation model designed to generate quality text from user inputs. Using Llama 3.1 effectively to generate guided outputs requires a structured input format— a prompt— to interact with the model. The clarity and the context the prompt provides are essential to succeeding model responses. The Llama 3.1 prompt comprises text sequences of tuned tokens and roles that are processed by the model. In Figure 1, we show a typical prompt we used for assessing query expansion. The `system` role sets the context to the model, the `user` conveys a retrieval query or statement that constitutes a textual question, and the `assistant` represents the response issued by the model.

Existing QA methods primarily focus on single data sources, either structured or unstructured. Recently, a growing interest in operating on questions that require reasoning of information from both tabular and raw text data sources simultaneously has gained traction— (Table-Text QA; Agarwal, Devaguptapu, and S 2025). In our paper, we evaluated QA distinctly over both table and text on (FeTaQA; Nan et al. 2022) and (AmazonQA; McAuley and Yang 2016; Gupta et al. 2019) datasets, respectively. FeTaQA is a collection of question-answer pairs from high quality Wikipedia tables, and its generative table question answering is formulated as an encoder-decoder learning problem. AmazonQA is rather user-review based and query expansion is commenced in the LLM by concatenating a top-ranked relevant review with the corresponding question.

2 Background

Query Expansion Broadly studied in the past several decades (Carpineto and Romano 2012), query expansion is

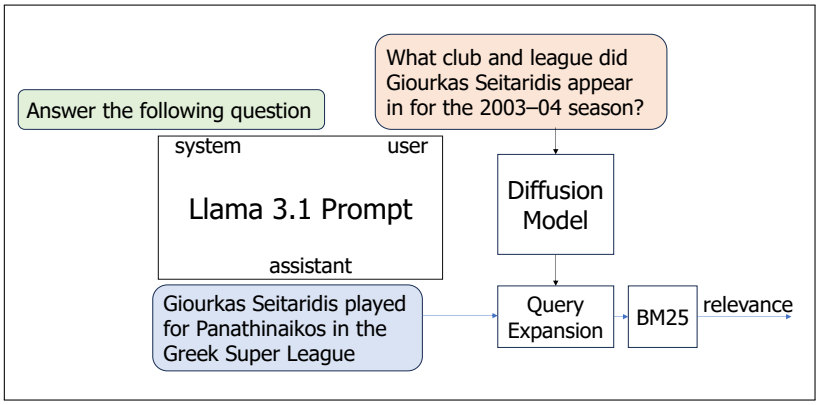

Figure 1: Overview of prompting the Llama 3.1 LLM for performing query expansion. Shown are the textual representations of the prompt roles, including the system, user, and the assistant. The initial query provided by the user is integrated in the forward phase of an LSTM-based diffusion model to yield a variety of sub-queries that are further concatenated with the language model response emitted by the assistant.

a foundational technique employed in information retrieval and question answering tasks, and performed either manually, automatically, or interactively. The method is used to improve neural search efficacy of sparse retrieval systems by rewriting the initial query with additional contextual terms based on pseudo-relevance feedback (PRF; Rocchio 1971; Robertson 1990) or top-ranked external knowledge sources. Conversely, document expansion in (Doc2query; Nogueira et al. 2019) augments the content of a document before indexing by training a sequence-to-sequence model to predict pseudo-queries based on documents, and subsequently adds generated pseudo-queries to the document index. In recent years, LLMs have shown to provide relevant information to guide retrieval systems. They demonstrate the effectiveness as query expansion models by generating pseudo-documents conditioned on few-shot prompts. The recent studies have also highlighted the emergence of prominent open-source LLMs, like Llama 3 (Meta 2024) that provide structured and well formatted prompts to better control the model output.

Embedding Diffusion Models Continuous diffusion models have been extremely successful in reasoning vision (Rombach et al. 2022) and audio (Sang-gil Lee et al. 2022) modalities, but their adaptation to text remained a challenge due to the inherent discrete nature of text. Diffusion models present a novel noising paradigm and a training objective other than token prediction that commonly found a language model. Diffusion models utilize a forward process to perturb the data with Gaussian noise, and a reverse denoising task to restore the data symmetrically. Recently, embedding diffusion models (Li et al. 2022; Gao et al. 2024) introduced an additional embedding step that converts discrete tokens into a pretrained learning representation. Adding Gaussian noise to the embeddings facilitated a fine-grained noising procedure to form a continuous diffusion process. A rounding step in the denoising phase turns predicted embeddings back to distinct tokens. In this paper, we explore the use of diffusion models operating in embedding space for query expansion.

3 Related Work

In his study, Claveau (2022) explored the use of text generation to automatically expand queries. As an early research to harness LLM power for the retrieval task, they used the GPT-2 model and showed that text generation is a highly effective approach to improve the performance of an IR system by a large margin with an average precision gain of 10% mAP@0.5. Wang, Yang, and Wei (2023) introduced Query2Doc for ameliorating quality of both sparse and dense retrieval systems. Query2Doc focuses on employing a version of GPT-3 model for generating passages related to the potential answers, aiming to alleviate the issue of word mismatch between a query and documents. Experimental results demonstrate that Query2Doc boosts the relevance performance of BM25 (Robertson and Zaragoza 2009) by 3 to 15 percent on canonical IR datasets, such as MS-MARCO and TREC DL. Jagerman et al. (2023) used Chain-of-Thought prompting (CoT; Wei et al. 2022b) for guiding query reformulation. CoT prompts instruct the language model to split its response gradually that leads to generating essential keywords for query expansion. Inviting reproducibility and openness of research, they solely experimented within the confines of the FLAN LLM family (Wei et al. 2022a), and observed higher quality of query expansion for denser models with 20B parameters. In their paper, Chen et al. (2024) addressed the application of query expansion to the open-domain question answering (OpenQA) task, and introduced a three-phase reasoning process for answer-oriented question expansion: a) The query is analyzed, b) Response-oriented expansions are generated, and c) A refinement step improves quality of query reformulation. Their framework in Llama 2 LLM outperforms state-of-the-art baselines in out-of-domain zero-shot scenarios. More recently, Seo and Lee (2025) offered QA-Expand that leverages LLMs to generate diverse question-answer pairs from an initial query. Their system rewrites the most informative pseudo-answers for effective query augmentation, and was shown to outperform state-of-the-art baselines by up to 13%.

To the extent of our knowledge, employing diffusion modeling to query expansion by combining LSTM, BERT, and Llama 3.1 models has not been explored in prior work.

4 Datasets

In our evaluation, we explored query expansion for QA over table using a BM25 retriever, contrasted with QA over text baseline for document relationships.

FeTaQA Table QA systems evaluate reasoning of query over tabular data. The recent free-form table QA dataset (FeTaQA; Nan et al. 2022), [1] frames generative table question answering as a problem of producing an answer a to a question q based on a table T and its metadata m, with the goal of constructing a table QA dataset $\{(q_i, a_i, T_i, m_i)\,|i = 1 \ldots n\}$ of a large number of instances n. Question-answer pairs are solicited from high quality Wikipedia tables and rather than short-span based, answers are free-form and long. Their end-to-end approach models the table QA task as a sequence-to-sequence learning problem by appending table linearization T to question q to form the source sequence, and projecting the free-form answer a as the target sequence. This process resembles retrieval-augmented generation (RAG; Lewis et al. 2020), by which the prompt is a string concatenation of the query and a knowledge base.

AmazonQA As a baseline QA over text we chose the large-scale review-based (AmazonQA; McAuley and Yang 2016; Gupta et al. 2019) dataset. [2] The dataset leverages an extensive community QA data and a rich volume of product reviews collected from product pages on Amazon.com, aiming to automatically learn what makes a review of a product relevant to a query. Undergoing information retrieval (IR) techniques, the top-k review snippets provided in the dataset were ranked for relevance to the query. Consistent with FeTaQA, the question q is prepended to the top review R for an augmented version of the input source to the LLM.

5 Method

In Figure 1, we provide an overview of generative query expansion in prompting Llama 3.1. Given the system role that sets the context to the model and the user who conveys the initial text query as LLM input, the expanded query q_e is the string concatenation of the text snippet representation for the corresponding user and assistant roles. This is expressed as a two-step reasoning process:

$$r = \text{generate}(\text{prompt}(s, \text{concat}(q, d))) \quad (1)$$

$$q_e = \text{concat}(\{q\} \times n, r), \quad (2)$$

where s sets the high-level context, q is the initial query, d the augmented knowledge data for retrieving the answer, and r the generated response, the answer, issued by the model. In the second step, the response r is string concatenated with the original query q. Often, in QA tasks the query q token-length is much shorter than the generated free-form answers, or for that matter the human curated responses. To balance the relative weights of the query and the corresponding answer, we follow common practice to increase the query term weights (Wang, Yang, and Wei 2023) by replicating the original query n times before concatenating it with either the model response for QA over table, or based on the score of expert voting in extractive QA over text. In the event of continuous generation, the equations above are generalized to perform the computation of q_e recursively. q_e is used as the new query for evaluating IR relevance or QA answer quality using either BM25 or embedding similarity approaches, respectively.

6 Experiments

We evaluated our methodology on two large-scale open-domain question answering datasets contrasting table QA with text QA. In our experiments, we used the instruction-tuned Llama 3.1 (Meta 2024) foundation model that renders a dense Transformer architecture of 8 Billion parameters. [3] We report results in either normalized BM25 scores or similarity measures over native Llama 3.1 embeddings. To account for uncertainty in our analysis, we also assess the level of agreement between our raters using the Kappa statistical measure. We ran inference locally and entirely on the CPU with up to four workers, while not exceeding 7.5GB of system memory. Our running time last about three minutes on average for each instance in any of FeTaQA and AmazonQA test splits. Training setup details for our experiments on the Diffusion-QE model are detailed in Paragraph 6.

Table QA FeTaQA is considered the first dataset for generative question answering over tables. FeTaQA contains 10,330 instances of which the test split comprises 2,003 question-answer pairs. Unlike span-based QA that mostly contain copies of short text spans from the immediate source, FeTaQA provides elaborate generative answers using T5 as an end-to-end model that integrates query and table understanding, logical reasoning, and language generation. In Table 1, we show core statistic distributions of FeTaQA. The free-form answers have a median of 18 tokens in length, and are grounded to the table. In contrast, questions have a smaller median of 12-token long, suggesting nonetheless a reasonable balanced dimensionality of question-answer pairs. Given the range of 35 and 52 tokens for question and answer length, respectively, the replication factor n of the original query q for computing the expanded query q_e is bound to a modest extent from 1 to 4.

Property	Range	Median	Mean
Question Length	35	12	12.5
Answer Length	52	18	19.7

Table 1: Core statistical distributions of FeTaQA dataset.

We used the rank-bm25 implementation for BM25 com-

[1] https://github.com/Yale-LILY/FeTaQA/tree/main/data

[2] https://github.com/amazonqa/amazonqa

[3] https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct

putation, [4] and chose the widely approved okapi scoring algorithm for a full text search query. BM25 ranking provides two parameters, $k1$ and b for tuning the relevance score calculation. $k1$ controls the scaling function between the term frequency of each matching term to the final relevance score of a document-query pair. $k1$ values are generally delimited between 0.0 to 3.0, with a default of 1.2. While b controls how the length of a document affects the relevance score. b values are in $[0, 1]$, with a default set to 0.75.

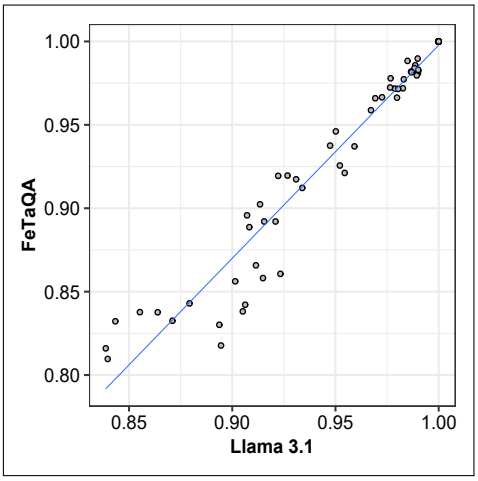

Figure 2: Scatter plot of BM25 scores for the relevance of each the Llama 3.1 and FeTaQA generated answers to the query.

In our evaluation, we rated the relevance of both the Llama 3.1 and the FeTaQA generated answers to a query, we further denote (q, r) and (q, a), respectively. In Figure 2, we present a scatter plot that shows the relation between BM25(q, r) and BM25(q, a) scoring pairs for 100 randomly selected instances from the FeTaQA test split. To improve clarity, we only depict the normalized BM25 scoring in [0.8,1.0]. We assessed the agreement between our two relevance raters by computing Cohen's Kappa measure (Cohen 1968). Similar to correlation coefficients, Kappa values range from –1 to +1, such that the higher the value of Kappa, the stronger the agreement. Kappa values for our two observers over the same 100 subjects are: $\kappa = 0.204$ indicating a fair agreement between the raters, $z = \kappa/\text{se}(\kappa) = 10.5$, where se is the standard error, and $p\text{-}value = 0$ implies that the agreement is statistically significant.

Property	Range	Median	Mean
Question Length	57	11	13.8
Answer Length	52	28	31.2
Review Length	778	72	70.4

Table 2: Core statistical distributions of AmazonQA dataset.

Text QA In our evaluation, we used the AmazonQA test split that contains 92,726 question-answer-review instances divided into 17 product categories. A question-context pair is considered answerable if the answer to the question is at least partially contained in the reviews. The answers provided by AmazonQA are extracted from user interaction in real-world scenarios and are either span-based or free-form. In our experiments, we sought after consistency with unbound answer generation by Llama 3.1 and only accepted answerable questions. Obtained by using BM25 scoring, AmazonQA provides query-relevant review-snippets that we apply as the knowledge base d for each question. In Table 2, we show core statistic distributions of AmazonQA. The median length of 11 and 28 tokens for a question and answer, respectively, fairly resembles the statistics we presented for FeTaQA. Whereas the review range close to 800 tokens is of considerable length, but still less than 4,096 tokens— the maximal input size for the 8B-parameter Llama 3.1 model we used.

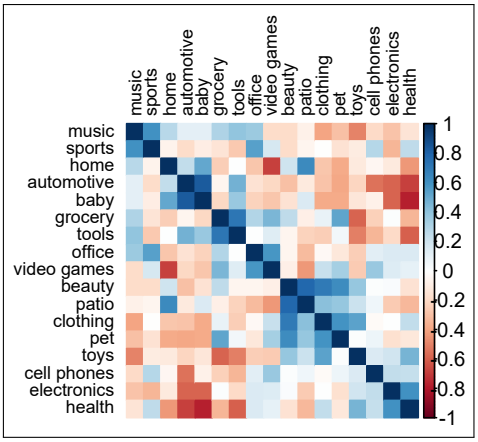

Figure 3: Correlogram of category relevance relations across AmazonQA product groups (red and blue colors render negative and positive relationships, respectively).

Besides textual description, documents are often interconnected with a certain type of relations (Xia et al. 2025). In addition to semantically similar query expansions for a given product, we are also interested in product relations. In Figure 3, we show a correlation plot across the 17 product categories spanned by AmazonQA. Applying hierarchical clustering, we drew all-pairs category correlations on similarity scores. We used cosine similarity measure between the generated answer and the human crafted answer voted by the review snippets, using the native Llama 3.1 embedding representation. Most prominent on the plot is a distinct product cluster composed of the highest positive relationships that is drawn symmetrically along the correlogram diagonal (dark blue). The cluster constitutes the product group $C =$ (beauty, patio, clothing, pet) with a Pearson correlation coefficient that ranges from 0.42 relatedness for pet-clothing to 0.79 for beauty-pet linkage. The relationship strength of product pairs we observed within the members of cluster C are exceptionally explicable.

Diffusion-QE In exploring a diffusion model for query expansion, we combined a BERT model pretrained on uncased

[4] https://pypi.org/project/rank-bm25/

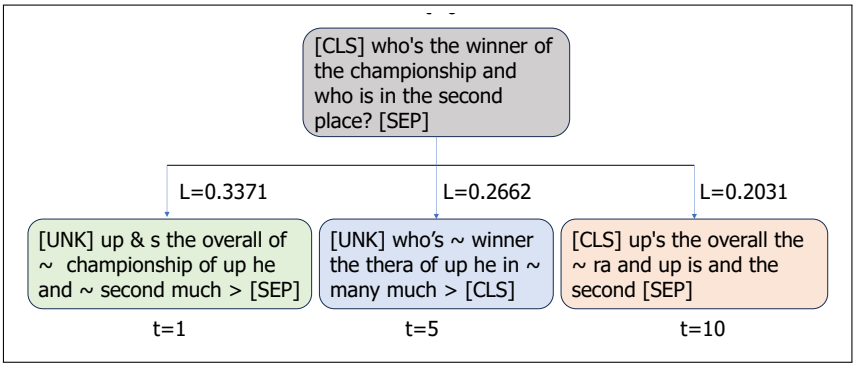

Figure 4: Overview of the diffusion forward phase: showing the original query q_0 at the top, followed by the generated query samples (q_1^d. q_5^d, q_{10}^d) that are impacted by increasing noise as a function of advancing the train timestep t. The loss L renders a gradual decline as time progresses.

English that feeds a Long Short-term Memory (LSTM) network with embeddings. [5] The merging of LSTM temporal feature extraction and the diffusion process of probabilistic modeling enhances the model predictive capacity. To train the diffusion model for query expansion, we defined a forward process that constructs the intermediate latent variable $q_{1:T}^d$, where q_t^d is a diffusion produced query at epoch t in either embeddings or textual representation, and T is the number of training epochs. At each training iteration, the forward process incrementally adds a Gaussian noise to the original unperturbed query q_0. In Figure 4, we highlight the forward diffusion phase, showing the original query and the ensuing continuous generative queries affected by the increased noise as the training session advances. We note that training the diffusion model for reversing the forward process and reconstructing the data is outside the scope of this paper. Our parameter setup for training includes the BERT default vocabulary size of 30,522 tokens, a batch size of 32, and embeddings and hidden state dimensions of 128 and 256, respectively. We chose using BERT over Llama 3.1 embeddings to considerably save both computation and memory resources. Our training session ran for ten epochs, using the Adam optimizer and a cross-entropy loss function.

The generated queries in the forward phase q_t^d are diverse and structured along a timeline that implies a pre-order by their relevance to the original query q_0. These sub-queries challenge the common practice for increasing the query term weights by replicating and concatenating the original query q_0 to balance the often longer token length of the model response (see Equation 2). Rather than an ad hoc basis that is unsupported by rigorous theory, we offer the concatenation of continuous sub-queries $\left(q_0; q_1^d; \cdots; q_T^d\right)$ to improve the sampling of the query space. By substituting the replication of the fixed original query with our sequence of temporally varied query content, the agreement between the two observed raters over 100 FeTaQA examples improved from fair to shy of moderate. On the other hand, the relationship scores on AmazonQA came close to the no-diffusion model. We attribute this behavior to the long review snippet that requires significantly more distinct sub-queries.

[5] 'https://huggingface.co/Contents/bert-base-uncased'

Noise Strategies We followed work that model text in the continuous embedding space and applied Gaussian noise uniformly to every token of the query (Li et al. 2022). Rather, more recently Chen et al. (2023) introduced a Masked-Diffusion language model and proposed adding soft-masked noise to different tokens in the input text with the intuition that more important words would be more perturbed. They defined the importance of words in a sentence based on word relevancy and entropy criteria, and follow a descending order of informativeness to apply noise proportionally. Compared to a unified Gaussian noise strategy their semantic accuracy performance improved by about six percentage points from 75.3 to 81.6. In our work, we explored applying random weighted noise to query tokens and observed a slight increase in the Kappa scores on FeTaQA. Although evaluated on a different dataset collected from the restaurant review domain, Masked-Diffusion LM renders a fairly moderate gain on random noise of close to 3.5 percentage points compared to a unified noise and much concurs with our behavior.

7 Conclusion

In this paper, we developed a framework for table and text question answering tasks to enhance the LLM query expansion generation. Using chain-of-thought prompting in Llama 3.1 for retrieval, our experiments on two large-scale QA datasets demonstrated a fair agreement between two generative raters and a sound document relationship across product categories, respectively. The approach we presented reduces reliance on human-provided answers and expert interventions, illustrating a sustainable method for enhancing query expansion generalization. To address the path dependency problem incurs in using pretrained LLMs on mature data, rather than repeating the question we considered to split the question into contextual chunks before generating the expansion. Our initial study that integrates the forward phase of a diffusion model for reasoning continuous text generation depicted compelling QE quality. A plausible research venue to further explore our work is fine-grain integration of QE in a wider scope of diffusion language models.

Limitations

We acknowledge several limitations in our paper. Our paper relies on open-source foundation models for evaluating query expansion, leaving closed-source LLMs outside the scope of this study, mainly due to affordability. Choosing Llama 3.1 as our baseline language model has seemingly hampered exploring other instruction-tuned LLMs to their full extent for assessing query expansion quality. However, Llama 3.1 has the advantage of incorporating a scalable prompt model to finely control LLM outputs. The process of query expansion in LLMs is known to incur latency for retrieving the local knowledge base, however, an LLM takes the same amount of computation for each generated token, thus the LLM performance is reasonably predictable. We note that models from the T5 family were fine-tuned on the FeTaQA train set, setting a competitive disadvantage to Llama 3.1. FeTaQA is categorized into 15 diverse topics, however, their distribution across small sample sizes hindered us from performing a tangible relational study comparable to AmazonQA.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful suggestions and feedback.

References

Agarwal, A.; Devaguptapu, C.; and S, G. 2025. Hybrid Graphs for Table-and-Text based Question Answering using LLMs. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 858–875. Albuquerque, New Mexico: Association for Computational Linguistics.

Carpineto, C.; and Romano, G. 2012. A Survey of Automatic Query Expansion in Information Retrieval. *ACM Computing Surveys*, 44(1).

Chen, J.; Zhang, A.; Li, M.; Smola, A.; and Yang, D. 2023. A Cheaper and Better Diffusion Language Model with Soft-Masked Noise. In Bouamor, H.; Pino, J.; and Bali, K., eds., *The Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4765–4775. Singapore: Association for Computational Linguistics.

Chen, X.; Chen, X.; He, B.; Wen, T.; and Sun, L. 2024. Analyze, Generate and Refine: Query Expansion with LLMs for Zero-Shot Open-Domain QA. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics (ACL)*, 11908–11922. Bangkok, Thailand: Association for Computational Linguistics.

Claveau, V. 2022. Neural text generation for query expansion in information retrieval. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 202–209. New York, NY, USA: Association for Computing Machinery. ISBN 9781450391153.

Cohen, J. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70: 213–220.

Gao, Z.; Guo, J.; Tan, X.; Zhu, Y.; Zhang, F.; Bian, J.; and Xu, L. 2024. Empowering Diffusion Models on the Embedding Space for Text Generation. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 4664–4683. Mexico City, Mexico: Association for Computational Linguistics.

Gupta, M.; Kulkarni, N.; Chanda, R.; Rayasam, A.; and Lipton, Z. C. 2019. AmazonQA: A Review-Based Question Answering Task. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI*, 4996–5002. International Joint Conferences on Artificial Intelligence Organization.

Jagerman, R.; Zhuang, H.; Qin, Z.; Wang, X.; and Bendersky, M. 2023. Query Expansion by Prompting Large Language Models. arXiv:2305.03653.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 9459–9474. Curran Associates, Inc.

Li, X.; Thickstun, J.; Gulrajani, I.; Liang, P. S.; and Hashimoto, T. B. 2022. Diffusion-LM Improves Controllable Text Generation. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 4328–4343. Curran Associates, Inc.

McAuley, J. J.; and Yang, A. 2016. Addressing Complex and Subjective Product-Related Queries with Customer Reviews. In Bourdeau, J.; Hendler, J.; Nkambou, R.; Horrocks, I.; and Zhao, B. Y., eds., *International Conference on World Wide Web (WWW)*, 625–635. Montreal, Canada: ACM.

Meta. 2024. The Llama 3 Herd of Models. Technical report, Meta. https://arxiv.org/abs/2407.21783.

Nan, L.; Hsieh, C.; Mao, Z.; Lin, X. V.; Verma, N.; Zhang, R.; Kryściński, W.; Schoelkopf, H.; Kong, R.; Tang, X.; Mutuma, M.; Rosand, B.; Trindade, I.; Bandaru, R.; Cunningham, J.; Xiong, C.; Radev, D.; and Radev, D. 2022. FeTaQA: Free-form Table Question Answering. *Transactions of the Association for Computational Linguistics (ACL)*, 10: 35–49.

Nogueira, R.; Yang, W.; Lin, J.; and Cho, K. 2019. Document Expansion by Query Prediction. arXiv:1904.08375.

Robertson, S. 1990. ON TERM SELECTION FOR QUERY EXPANSION. *Journal of Documentation*, 46(4): 359–364.

Robertson, S.; and Zaragoza, H. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval*, 3(4): 333–389.

Rocchio, J. J. 1971. Relevance feedback in information retrieval. *SMART retrieval system: experiments in automatic document processing*, 313–323.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Sang-gil Lee; Kim, H.; Shin, C.; Tan, X.; Liu, C.; Meng, Q.; Qin, T.; Chen, W.; Yoon, S.; and Liu, T.-Y. 2022. PriorGrad: Improving Conditional Denoising Diffusion Models with Data-Dependent Adaptive Prior. In *International Conference on Learning Representations*.

Seo, W.; and Lee, S. 2025. QA-Expand: Multi-Question Answer Generation for Enhanced Query Expansion in Information Retrieval. arXiv:2502.08557.

Wang, L.; Yang, N.; and Wei, F. 2023. Query2doc: Query Expansion with Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 9414–9423. Singapore: Association for Computational Linguistics (ACL).

Wei, J.; Bosma, M.; Zhao, V.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2022a. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; ichter, b.; Xia, F.; Chi, E.; Le, Q. V.; and Zhou, D. 2022b. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems (NIPS)*, volume 35, 24824–24837. Curran Associates, Inc.

Xia, Y.; Wu, J.; Kim, S.; Yu, T.; Rossi, R. A.; Wang, H.; and McAuley, J. 2025. Knowledge-Aware Query Expansion with Large Language Models for Textual and Relational Retrieval. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 4275–4286. Albuquerque, New Mexico: Association for Computational Linguistics.