

Query Suggestion for Retrieval-Augmented Generation via Dynamic In-Context Learning

Fabian Spaeh

f.spaeh@celonis.com

Tianyi Chen*

ctianyi7@gmail.com

Chen-Hao Chiang

c.chiang@celonis.com

Bin Shen

b.shen@celonis.com

January 14, 2026

Abstract

Retrieval-augmented generation with tool-calling agents (agentic RAG) has become increasingly powerful in understanding, processing, and responding to user queries. However, the scope of the grounding knowledge is limited and asking questions that exceed this scope may lead to issues like hallucination. While guardrail frameworks aim to block out-of-scope questions (Rodriguez et al., 2024), no research has investigated the question of suggesting answerable queries in order to complete the user interaction.

In this paper, we initiate the study of query suggestion for agentic RAG. We consider the setting where user questions are not answerable, and the suggested queries should be similar to aid the user interaction. Such scenarios are frequent for tool-calling LLMs as communicating the restrictions of the tools or the underlying datasets to the user is difficult, and adding query suggestions enhances the interaction with the RAG agent. As opposed to traditional settings for query recommendations such as in search engines, ensuring that the suggested queries are answerable is a major challenge due to the RAG’s multi-step workflow that demands a nuanced understanding of the RAG as a whole, which the executing LLM lacks. As such, we introduce robust dynamic few-shot learning which retrieves examples from relevant workflows. We show that our system can be self-learned, for instance on prior user queries, and is therefore easily applicable in practice. We evaluate our approach on three benchmark datasets based on two unlabeled question datasets collected from real-world user queries. Experiments on real-world datasets confirm that our method produces more relevant and answerable suggestions, outperforming few-shot and retrieval-only baselines, and thus enable safer, more effective user interaction with agentic RAG.

1 Introduction

Retrieval-augmented generation (RAG) allows large language models (LLMs) to interact and manipulate data from external sources via tool calls (Lewis et al., 2020). RAG circumvents well-known limitations of LLMs such as the inability to expand their memory after training, carrying out exact computations, or their lack of domain-specific knowledge (Naveed et al., 2023; Tänzer et al., 2022; Brown et al., 2020). Even though breakthroughs in LLMs that facilitate RAG are recent, they are already deployed and enjoy increasing popularity (Patil et al., 2023; Zhang et al., 2024; Patil et al., 2024). Many LLMs now natively support external tool-calling functionality, simplifying the implementation of RAG agents. In our work, we consider RAG agents that go beyond free-text retrieval and support arbitrary tool calls, which is why they are also referred to as “agentic.” Such a system makes progress towards an answer via a workflow that is dynamically decided by the LLM. Such a workflow can include multiple steps of tool calls and reasoning, before a final response is delivered to the user.

*Work done while at Celonis, Inc.

Agentic RAG lowers the barrier of entry for interacting with the underlying data. On the flip side, this means that typical users will be unaware of the underlying tool calls and data representation. This introduces challenges in the RAG-user interaction and queries can often fail: For instance, users may ask for information that is not available, e.g., required information is not present in a database. Queries may also fail due to technical reasons, e.g., when a user wants to search in a column that is not indexed. To alleviate such issues, we propose query suggestion for agentic RAG. Specifically, we aim to suggest queries to the user that fulfill two goals:

1. The suggested query should be semantically similar to the user query and carry the same intent as the original query.
2. The suggested query should be answerable by the RAG.

To the best of our knowledge, this problem is novel in the context of agentic RAG. The closest to our problem is the work of [Tayal and Tyagi \(2024\)](#), which considers query suggestion for RAG for free-text retrieval. However, since this does not involve an arbitrary agentic workflow, it is similar to traditional work on query suggestion. Compared to this, we face additional challenges in our setup on agentic RAG: Queries can pertain to values that appear in any source (e.g., database) and the number of queries is much larger, because the agentic RAG can execute an arbitrary set of workflows. Understanding whether a question can be translated into an executable workflow by the agentic RAG is difficult, mainly due to the disconnect between the LLM, the available tools, and the underlying data. A priori, the LLM does not know the available data and executing a workflow successfully requires nuanced understanding of the available tools which is difficult to deliver via prompt instructions. Furthermore, in order to produce a valid response, RAG agents may undergo several layers of self-reflection ([Renze and Guven, 2024](#)) and confidence estimation ([Li et al., 2024](#)). As these are executed on higher level, understanding the outcome of a query without executing it becomes challenging.

1.1 Our Contributions and Techniques

We are the first to study the problem of query suggestion for agentic RAG and make the following contributions towards this problem.

1. We propose a query suggestion method based on few-shot learning. To enhance the LLM’s understanding about query answerability, we propose robust dynamic few-shot learning, which is a novel method to retrieve relevant examples and automatically combat hallucination. We facilitate retrieval through a templating approach, which can be understood as reducing a query to its workflow.
2. We show how training examples can be self-learned, which makes our approach immediately applicable. We use the RAG itself to answer and label queries, and we thus do not rely on large sets of labeled data. As such, our approach is practically applicable and can learn from prior user queries in an unsupervised way.
3. We experimentally evaluate our approach on three real-world benchmark datasets, where we consistently outperform static few-shot learning as a baseline and a retrieval-only approach. This also shows that static few-shot learning without the dynamic retrieval of examples does not lead to effective query suggestion, implying that a mere description of tools does not suffice for the complex task of query suggestion for agentic RAG.

On a conceptual level, we introduce a distinction between two kinds of unanswerable queries: Queries without a workflow to answer them, and queries for which the called tools cannot find the needed information in the underlying data.

1.2 Related Work

Query Suggestion Query suggestion aims to identify a follow-up query that the user may ask ([Cao et al., 2008; Mei et al., 2008; Baeza-Yates et al., 2004](#)). [Tayal and Tyagi \(2024\)](#) are the first to study query suggestion. However, their work does not apply to agentic workflows, but is instead tailored to free-text

retrieval. Agentic RAG is more challenging and requires a more sophisticated retrieval scheme, as it is much harder to guarantee answerability due to the multi-step workflow. Furthermore, we use a form of transfer learning which allows our approach to be entirely self-learning, while the method of Tayal and Tyagi (2024) relies on hand-crafted training examples. To the best of our knowledge, this is the only prior work on query suggestion for RAG, and it is not in our tool-calling setting.

Recent work on traditional query suggestions for search engines also suggests the use of LLMs (Mustar et al., 2022; Baek et al., 2024). This line of work is on general query suggestion, with no focus on similarity and answerability in the context of a RAG agent. Perhaps the closest work is on task-based search, where the query suggestion aims to pick up on a user’s intended task and tries to suggest queries that lead to the completion of the task (Ding et al., 2018; Garigliotti and Balog, 2017). This line of work aims to suggest queries that are related but also aid to achieve the user’s goals. None of these works consider query suggestion for RAG. Query suggestion is also a popular tool for enhancing search for retrieval systems, in which context it is referred to as query expansion. The idea is that multiple similar queries can enhance the retrieval if it is based on nearest neighbor search on embeddings (Ooi et al., 2015; Carpineto and Romano, 2012). Such systems have been employed for RAG, but only for free-text retrieval, i.e., without tool-calling agents (Ma et al., 2023; Chan et al., 2024).

LLM Guardrails LLM guardrails limit the scope of a language model agent to a dedicated purpose or a pre-defined interaction path. In the context of RAG, the agent should only answer questions that pertain to its dedicated task and are within the abilities of its tools. Guardrails typically intercept communication with the LLM, and allow to program constraints and conversation flows via an intermediate layer (Rebedea et al., 2023; Rajpal, 2023). Inan et al. (2023) introduce a dataset and classifier that detects different kinds of harmful LLM inputs or outputs. For a thorough discussion, we refer the reader to (Dong et al., 2024).

Query Disambiguation Another related line of research is on question disambiguation, where the goal is to respond with questions that clarify the original query. The goal is therefore similar to our work, as it is triggered from queries that are unclear and cannot be answered directly. Large language models have been used to generate such disambiguating questions for information retrieval (Zamani et al., 2020), for agents that answer open domain questions (Kuhn et al., 2022; Aliannejadi et al., 2020; Kim et al., 2023). This line of research does not extend to agentic RAG.

Organization In Section 2, we define three answerability categories and give a high-level overview of our approach. Section 3 details our dynamic few-shot learning approach and templating. In Section 4, we show how to use the RAG agent to automatically label example queries with their answerability. We follow up with an experimental evaluation in Section 5, where we also detail the deployment.

2 Query Suggestion for RAG

In order to guide the query suggestion, we need to make the following classification about the *answerability* of a query. On a high level, we differentiate between queries that are not answerable due to a limitation in the agent configuration or available tools, or due to limitations in the underlying data. We use the following three categories to classify queries.

1. *No workflow*: The RAG agent cannot answer the question as it cannot be transformed into an executable workflow given the RAG’s purpose and tools. This can be because the question is out of scope, or the RAG agent does not have access to the necessary tools to compute a solution.
2. *No knowledge*: The RAG agent understands the question, and is able to translate the question into a series of tool calls to answer it. However, there is no data pertaining to the specific values in the query. This appears whenever a tool connecting to external knowledge responds with an empty or invalid result, which makes the final RAG response not meaningful. As opposed to the first category, this stems from limitations in the underlying data.

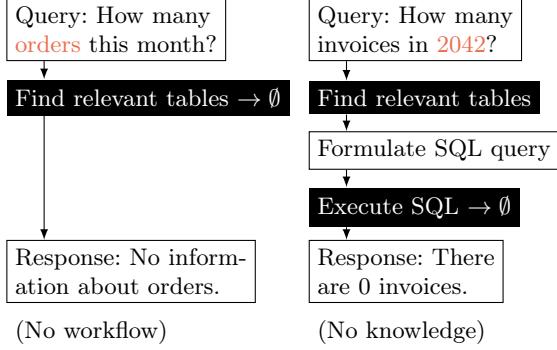


Figure 1: Two queries that are not answered. Black boxes denote a tool execution. (Left) the RAG agent does not have access to information about orders, i.e. it is out of scope and there is no workflow that will successfully answer the query. (Right) The query fails since the final tool call returns an empty response due to a typo in the year, indicating that there is no underlying knowledge about the year 2042.

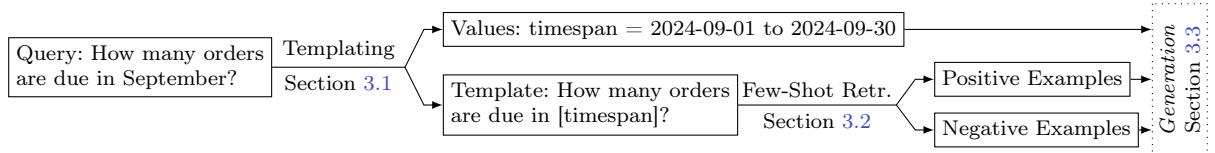


Figure 2: Dynamic Few-Shot Learning. Positive examples are labeled queries that are answerable; negative examples are queries for which there is no workflow.

3. *Answerable*: The RAG agent understood the question and delivered a meaningful answer. In this case, the user is satisfied with the answer and not interested in similar questions but would want to end the conversation or move on to different topics, which are not the focus of this work.

We showcase two examples to highlight the differences between *no workflow* and *no knowledge* in Figure 1. Note that this distinction is not clear-cut as a dynamic workflow can depend on the results of prior tool-calls. Furthermore, some queries demand complex workflows that the LLM may not be able to deduce, so even though there is a possible workflow, it cannot be found. For simplicity, we will ignore such issues as they are not in the focus of our work.

Our Approach The goal of our work is to suggest similar queries to user questions which fall into the first two answerability categories, and for which the user would want to follow up with a corrected query. In our approach, we first focus on the workflow of a query and ignore the specific values that appear in it. If a query is not answerable because there is no workflow to execute it, we then generate similar queries that can be executed. Here, we explicitly ignore values in the queries via templating, which we explain in the next section. The result is a query that can be answered if we put in sensible values that avoid issues with the data. In a second step, we therefore impute values into the query that make it fully answerable, while trying to stay similar to the original query.

Our basic approach to generate an answerable query builds on few-shot learning (Song et al., 2023), a methodology from meta-learning that avoids retraining the language model by providing only a few example queries from the training distribution. Furthermore, the examples we provide are not pairs of user queries and query suggestions. Instead, we transfer-learn by providing few-shot examples of answerable and unanswerable queries, alongside an explanation for the answerability. LLMs have shown great ability to utilize instructions and a few examples to generalize to unseen data (Wang et al., 2022), and this approach is also substantially cheaper than retraining the model. For our query suggestion problem, we use multiple examples that are each annotated with an explanation, to enhance the RAGs understanding of answerable queries and executable workflows. Specifically, we provide positive and negative examples. Positive examples are queries that are classified as answerable, and negative examples are queries that are not answerable because they have no

workflow. We ignore queries for which there is no knowledge since we cannot say with certainty whether the right combination of values may make them answerable. We enhance positive and negative example queries by providing an explanation why the RAG agent was able or unable to respond. This allows the LLM to understand tools and data better, in order to generalize better to unseen examples.

Since the LLM is trained on natural language, it can easily generate similar queries. However, discerning possible workflows is difficult and the answerability of suggested queries is therefore often low even with static few-shot examples, as we show experimentally in Section 5. To overcome this, we want to provide a much larger set of positive and negative example queries, allowing the LLM to infer from many examples that capture the nuances of the set of answerable queries better. However, the number of examples we can provide is tightly limited by the LLM’s context size. We propose a solution that dynamically retrieves the most relevant positive and negative example queries, which we provide as few-shot examples to the LLM. Selecting the right few-shot examples is critical for in-context learning. For instance, Agrawal et al. (2023) analyze the performance in translating a sentence, while comparing the n -gram overlap of the sentence with the few-shot examples. Our methodology also avoids retraining the model, which is beneficial as retraining can be expensive and time-consuming. Furthermore, in Section 4, we show how to use the RAG to automatically label training examples and self-learn.

3 Dynamic Few-Shot Learning

For this section, we assume access to a large set of training examples which are labeled with their answerability category and an explanation for why the category is chosen. Our goal is to identify relevant example queries as few-shot examples and generate query suggestions based on this. In the first step, we map a query to a workflow, which we achieve via a templating approach. In the second step, we show how this enables us to retrieve relevant few-shot examples in a robust manner. Finally, we show how to use these few-shot examples to generate query suggestions. A complete overview is given in Figure 2.

3.1 Question Templating

Text embeddings are a standard approach for semantic retrieval and have shown success in recommender systems, providing high recall (Abdullahi et al., 2024). Embeddings allow for fast evaluation of semantic similarity and nearest-neighbor search, for instance via the cosine similarity.¹ Generating queries that have an executable workflow requires relevant few-shot examples. The relevance is therefore not dependent on the specific values that appear in a query. Instead, we want to retrieve queries that demand similar workflows, some of which can be executed (positive example) and some of which not (negative example). Whether an entity is a value is agent-specific, since it depends on the available tools and the underlying data. Hence, retraining an embedding model for each agent configuration is expensive and requires large amounts of training examples. We therefore follow a different approach that avoids fine-tuning and is able to robustly map a query to a workflow, by tasking an LLM to mask all entity values.

Our aim is to replace all entity values by their entity name. Specifically, we want to replace all entity values that will become arguments for a tool call, such that a template simply corresponds to the workflow that answers the query. Here is a simple example of our templating, in which we introduce a mask for the timespan:

“How many invoices were processed in September 2021?”

→ “How many invoices were processed in [timespan]?”

Note that the RAG workflow stays the same for all timespans. As templates do not pertain to specific values, they cannot have data issues and are therefore either answerable or not answerable.

Templating is an easy task for LLMs, but we need each tool to describe its arguments, along with the type and a few example values. We provide this and instruct the LLM to replace all entity values with their entity names. We facilitate this with an instruction prompt and few-shot examples.

¹The cosine similarity of two embedding vectors $e, e' \in \mathbb{R}^n$ is equal to their normalized inner product $\langle e, e' \rangle / (\|e\|_2 \cdot \|e'\|_2)$. For this paper, we will assume that embedding vectors are already normalized, i.e. $\|e\|_2 = \|e'\|_2 = 1$.

3.2 Dynamic Few-Shot Retrieval

Our goal is to provide relevant few-shot examples to the LLM that cover the possible workflows via template queries. However, even though the space of template queries is much lower than that of concrete queries, the number of potential workflows is still large. This introduces two issues: We cannot enumerate all possible template questions, and even if we had access to a complete list of template questions, we could not possibly provide all of them to the LLM as this would quickly exceed the LLM’s context size. We avoid this problem by retrieving few-shot examples dynamically, via an embedding of the template query. This ensures that few-shot examples are always relevant.

An immediate problem with the previous approach is that (almost) identical queries result in different answers. One reason for this lies in faulty outputs as a result of hallucination in complex multi-step queries (Huang et al., 2023). Such responses, which we refer to as faulty, may easily poison the examples. To mitigate the effect of faulty responses, we make our retrieval approach robust by approximating a local majority vote among the retrieved examples.

We detail our approach in Algorithm 1. On a high level, we first retrieve few-shot examples that are relevant to the user query. We then cluster the retrieved queries by their similarity and outputs one query per cluster. The answerability of this query is decided based on the majority. E.g., if more than half of the queries in a cluster are considered not answerable, we use this as a label. Our concrete implementation of this strategy is as follows. We denote with \hat{e} the embedding of the templated user query. We use two thresholds θ_{sim} and θ_{div} . The former is used to decide the addition of few-shot examples when comparing with embedding \hat{e} , the latter to decide the removal of few-shot examples that are too similar to each other. We keep a counter for each example and iterate through all examples in arbitrary order. Instead of directly outputting an example i that is similar to the user query, we search other examples j with similarity that exceeds $\langle e_i, e_j \rangle \geq \theta_{\text{div}}$. If such an example exists, we instead increase or decrease the counter for j if the answerability of i matches the answerability of j or not, respectively. The rationale behind this is to ensure that the confidence on the answerability of each few-shot example is high, and we use the described procedure to approximate a majority vote.

3.3 Generation

Based on the dynamically retrieved few-shot examples, we task the LLM with generating one (or multiple) templates that serve as answerable templates for our query suggestion. Each template will be converted into a suggestion, so we obtain multiple suggested queries. An example of such a prompt is detailed in Figure 4.

To generate a query suggestion, we need to replace the masks of the generated templates with actual values, making the query fully answerable without data issues. For this, we provide the full history of tool-calls, reasoning steps, and the final response to the LLM. We task the LLM to find values for each mask in this order:

1. Consider the values of the original query that are used as tool-call arguments. If a data issue occurred because of a value, the value is ignored. Otherwise, use the same value.
2. Whenever a value is problematic, a tool may already provide good alternative values in its response. If this is the case, use the provided value.
3. If the value is problematic and the tool does not provide alternatives, pick example values from the tools.

The steps ensure that the values picked are as close as possible to the original query, while keeping the suggestion answerable. Note that this requires each tool to provide good example values from the data it accesses. Furthermore, programming the tool to suggest alternative values can improve the suggestion quality greatly.

4 Self-Learning

So far, we assumed the existence of a large set of labeled queries. However, to make our dynamic few-shot approach practically applicable, we now show that we can use the RAG agent itself to label queries with

```

RetrieveExamples( $\hat{e}$ ,  $\theta_{\text{sim}}$ ,  $\theta_{\text{div}}$ )
   $E \leftarrow \{(e_i, \text{ans}_i, \text{tmpl}_i) \in D : \langle \hat{e}, e_i \rangle \geq \theta_{\text{sim}}\}$ 
  Let  $n = |E|$ 
   $c_1 \leftarrow 1$ 
   $c_2, \dots, c_n \leftarrow 0$ 
  for  $i = 2, \dots, n$  do
    Let  $j^* \leftarrow \arg \max_{1 \leq j \leq n} \text{with } c_j > 0 \langle e_i, e_j \rangle$ 
    if  $\langle e_i, e_{j^*} \rangle \geq \theta_{\text{div}}$  then
      if  $\text{ans}_i = \text{ans}_{j^*}$  then
         $| c_{j^*} \leftarrow c_{j^*} + 1$ 
      else
         $| c_{j^*} \leftarrow c_{j^*} - 1$ 
      end
    else
       $| c_i \leftarrow 1$ 
    end
  end
   $E^+ \leftarrow \{\text{tmpl}_i : c_i > 0 \text{ and } \text{ans}_i = \text{answerable}\}$ 
   $E^- \leftarrow \{\text{tmpl}_i : c_i > 0 \text{ and } \text{ans}_i = \text{not answerable}\}$ 
  return  $E^+, E^-$ 
end

```

Algorithm 1: Robust retrieval of dynamic few-shot examples. Example templates tmpl_i are stored along their normalized embedding vector e_i and the answerability $\text{ans}_i \in \{\text{answerable}, \text{not answerable}\}$. We measure all distances between template embeddings e_i and e_j using the cosine-similarity $\langle e_i, e_j \rangle$. Additionally, we limit the positive examples E^+ to the 5 templates which are most similar to \hat{e} , and do the same with the negative examples E^- .

their answerability category and provide an explanation for the categorization. As such, our overall approach requires only a set of unlabeled queries. We can learn directly on user queries, which additionally avoids a skew between queries from the training data and real-world queries. Figure 3 shows our overall self-learning approach and we show the space of self-learned examples in Appendix A.

There are works that use similar ideas where the LLM evaluates its own responses and this evaluation is used to inform future decisions. For instance, Madaan et al. (2023) suggest to iteratively improve outputs for a user question based on LLM feedback. Shinn et al. (2023) enable reinforcement learning through a reflection module which generates response feedback via an LLM. However, combining LLM feedback with our few-shot retrieval approach extends the available feedback beyond the LLMs context size, which is a challenge for prior work.

4.1 Response Evaluation

It is difficult for an LLM to judge the answerability of a query up front as the visible information about the tools and underlying data is limited. However, it is possible to determine the answerability in hindsight after the RAG agent executed the query.

As apparent from Figure 3, we do this by showing the user query, the full chain of tool calls and tool responses, reasoning steps, and the final RAG response to the LLM. As such, this is akin to self-reflection (Rebedea et al., 2023) but we do not ask the LLM to judge the correctness of the answer. Instead, we task the LLM with classifying the answerability of the query. We facilitate this via a prompt that contains the instruction to evaluate the answerability, along few-shot examples for each answerability category with an explanation why the category applies. We also task the LLM to output an explanation why the specific category is chosen.

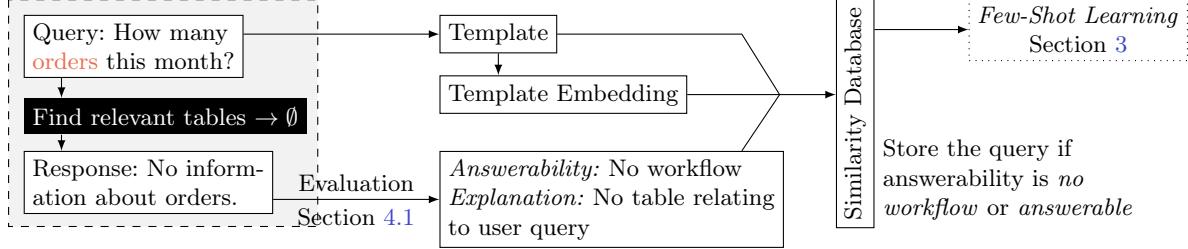


Figure 3: Simultaneous learning and query suggestion. The LLM evaluates the answerability of the query based on the chain of tool-calls and response produced by the RAG agent. The evaluation is stored in the similarity database using the embedding vector of the templated query. At the same time, we use the template query for dynamic retrieval of few-shot examples.

5 Experimental Evaluation

We now provide an extensive evaluation of our approach on three real-world benchmark datasets. Note that since we are the first that address the problem of query suggestion for agentic RAG, there is no prior work. We thus compare our approach against two natural baselines. Our results show that our approach outperforms both baselines with only a few thousand unlabeled training examples. We label the training examples them with their answerability as described in Section 4.

5.1 Experimental Setup

Retrieval-Augmented Generation We use a proprietary query-generating copilot that executes arbitrary natural language queries via dynamic workflows. It operates on top of a knowledge base akin to a traditional relational database. This RAG agent is provided with the database schema, and queries to retrieve and display content from the knowledge base. It offers functionality for ad-hoc calculations via a tool that executes Python code. The RAG agent contains a self-reflection module that evaluates the RAG response and triggers a re-evaluation if a mistake in the workflow is detected. It is built with GPT-4o ([OpenAI, 2023](#)) in the 2024-08-06 version. We use three different RAG agents. Two agents are tasked with invoice processing, whereas the first (INVOICESNOPYTHON) does not have access to the Python tool and the second (INVOICESPYTHON) does have access to the Python tool and can therefore carry out ad-hoc calculations. The abilities of both agents therefore differ starkly. The third agent is tasked with order management (ORDERS).

Datasets We obtain two datasets from real-world user logs, recorded over a period of two weeks. The questions are from related tasks, but for different RAG agents which also have different configurations. This results in ≈ 2000 questions for each agent. Because of this, most questions are related but not unanswerable, as the RAG agents differ. We show dataset statistics in Figure 5. For our experiments, we report values from a five-fold cross-validation.

Query Recommendation In our evaluation, we use two natural baselines alongside our approach.

1. *Static few-shot*: Query suggestions are generated based on instructions and a static set of few-shot examples, which do not depend on the input query.
2. *Retrieval-only*: Query suggestions are obtained directly from the queries retrieved by our robust retrieval approach as detailed in Section 3.2 where only output the most similar positive example. We merely impute values into the retrieved template as in Section 3.3, and task the LLM not to change the template.
3. *Dynamic few-shot*: Our method that retrieves examples dynamically and uses them as few-shot examples for the query suggestion generation, as described in Section 2. We provide at most 5 positive and 5 negative examples.

Instructions: You are an agent that suggests queries to questions that were not answered. The query “What is the average of days paid late for [timespan]” was not answered. . . .

Positive examples:

- Template: What are common paths for late payments?
Explanation: This was answered because ...
- Template: How to improve average days paid late during [timespan]?
Explanation: This was answered because ...

Negative examples:

- Template: Show average days paid late for [company code].
Explanation: Response did not provide specific information.
- Template: What is the average invoice value in [timespan]?
Explanation: The response provided a Python code snippet which does not accurately answer the question.

Task: Generate an answerable template query that is similar

Figure 4: Prompt for the generation of an answerable template query. For brevity, we do not show the full prompt and explanations for answerable example queries. The explanation for the second negative example gives a strong hint that Python code will not be executed, so ad-hoc calculations are not allowed for this instance. As a result, we obtain the suggestion “What is the total number of invoices paid late in [timespan]?” which can indeed be answered.

	Name	# total	no workflow	no knowledge	answerable
INVOICESNoPYTHON	2029	1413	225	391	
INVOICESPYTHON	2029	1295	246	488	
ORDERS	2766	384	1322	1096	

Figure 5: Dataset statistics. We show the number of queries along with their classification, as described in Section 4.1.

Metrics We evaluate both the answerability and the similarity to the original query. In order to obtain a score for the answerability (i.e. not answerable, no knowledge, or fully answerable), we task the LLM to evaluate the Copilot’s response to the original query along with all the intermediate tool calls as described in Section 4.1. In order to measure the semantic similarity between the user query and the suggestion, we use the cosine-similarity of the embedding vectors. This is commonly understood as a good proxy for the semantic similarity (Freestone and Santu, 2024; Kim et al., 2022; Mathur et al., 2019). For simplicity, we only evaluate answerability and similarity on the first query suggestion.

Implementation Details We use a few practical optimizations to speed up the query suggestion. To reduce the number of generated tokens, when templating a query, we ask the LLM to provide the list of entity names and values that appear in a query, instead of asking to generate the masked sentence. We also combine template generation and value imputation. We ask the LLM to go through both steps but not to output the intermediate templates. For the generation of query suggestions, we also use GPT-4o (OpenAI, 2023). For template embeddings, we use text-embedding-3-small (OpenAI, 2024).

5.2 Results

Answerability and Similarity We show results for the three benchmark datasets in Figure 6. This shows that our dynamic few-shot method is able to outperform the two baselines static few-shot learning and the retrieval-only approach. As described in the introduction, we observe that static few-shot learning has a high

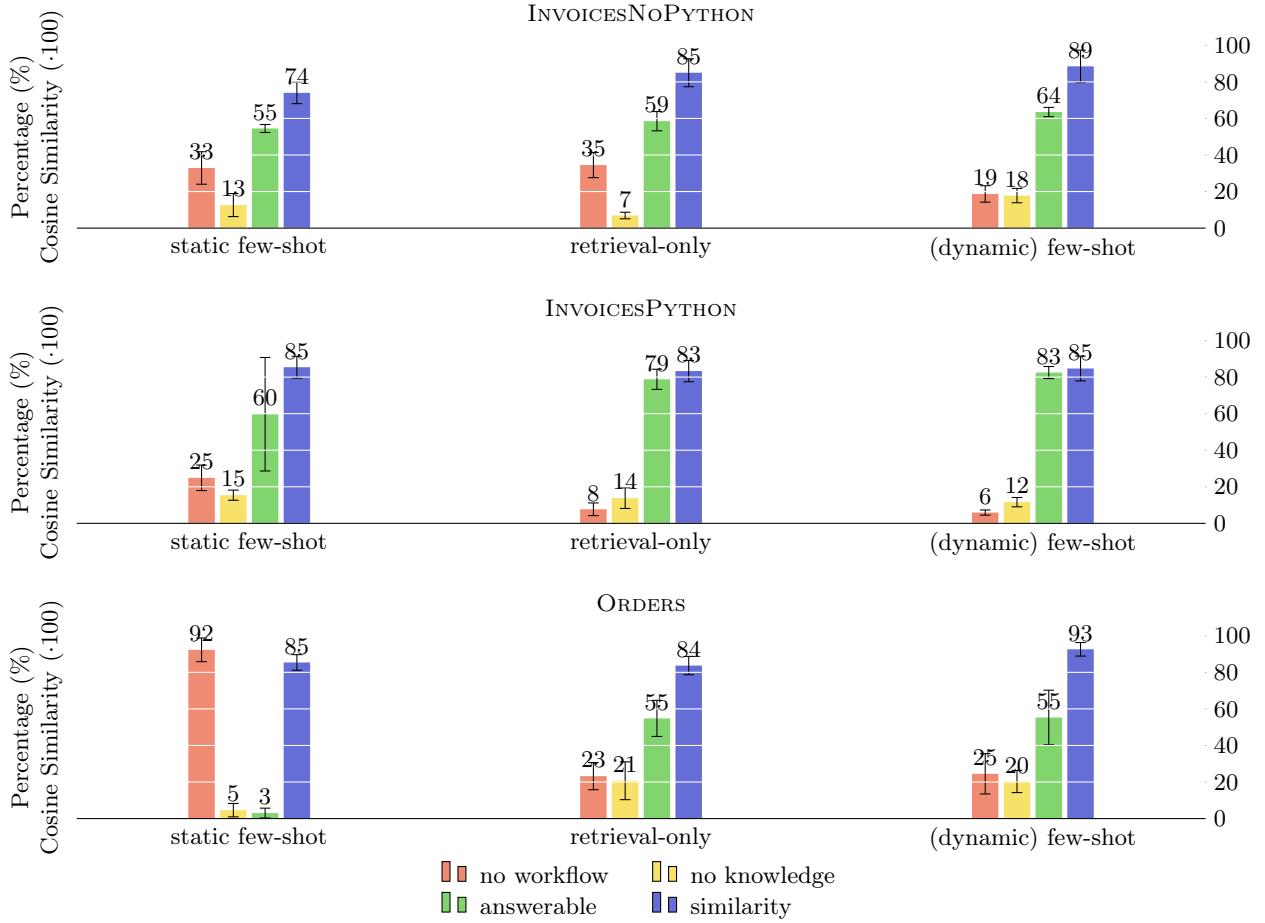


Figure 6: Dynamic few-shot learning produces significantly more answerable and relevant queries across all datasets. We group suggested queries into three answerability categories using the approach of Section 4.1 and show the percentage of queries in each category. Additionally, we evaluate the cosine similarity between the user query and the suggestion. We report mean and standard deviation across five runs.

trade-off between similarity and answerability. For instance, on INVOICESPYTHON, the similarity matches that of the dynamic few-shot learning approach, but the answerability is worse. This is particularly pronounced on the ORDERS dataset, where we can see that the LLM is able to generate only few answerable queries from the instructions and few-shot examples without feedback from the tools. The retrieval-only approach generally improves upon the static few-shot learning. An improvement in answerability is expected, as suggested queries are all classified as answerable. However, on our three benchmark datasets we observe that the retrieval-only suggestions also have high similarity to the user queries, suggesting that the distribution of user queries is sufficiently narrow. With dynamic few-shot learning, we further strictly improve in terms of similarity and answerability. Surprisingly, dynamic few-shot learning surpasses the answerability of the the retrieval-only approach. The latter seems to be a current barrier due to issues with hallucination, and many queries are challenging such that the RAG agent cannot consistently answer them. A reason why dynamic few-shot learning can nonetheless outperform the retrieval-only approach is that after seeing relevant positive and negative few-shot examples, the LLM decides on more conservative questions which can be answered with higher certainty.

Answerability-Similarity Tradeoff We show the detailed trade-off between answerability and similarity for individual suggested queries in Figure 9 in the appendix. This supports the general observation from

Figure 6 that the answerability and similarity of our dynamic few-shot learning approach dominates that of static few-shot learning and the retrieval-only approach. We also see that answerability of the retrieval-only approach degrades with increasing similarity. Dynamic few-shot learning exhibits little to no tradeoff, suggesting that it generalizes to unseen training examples. The answerability remains stable for most ranges of similarity, except for a drop-off for query suggestions that are very dissimilar.

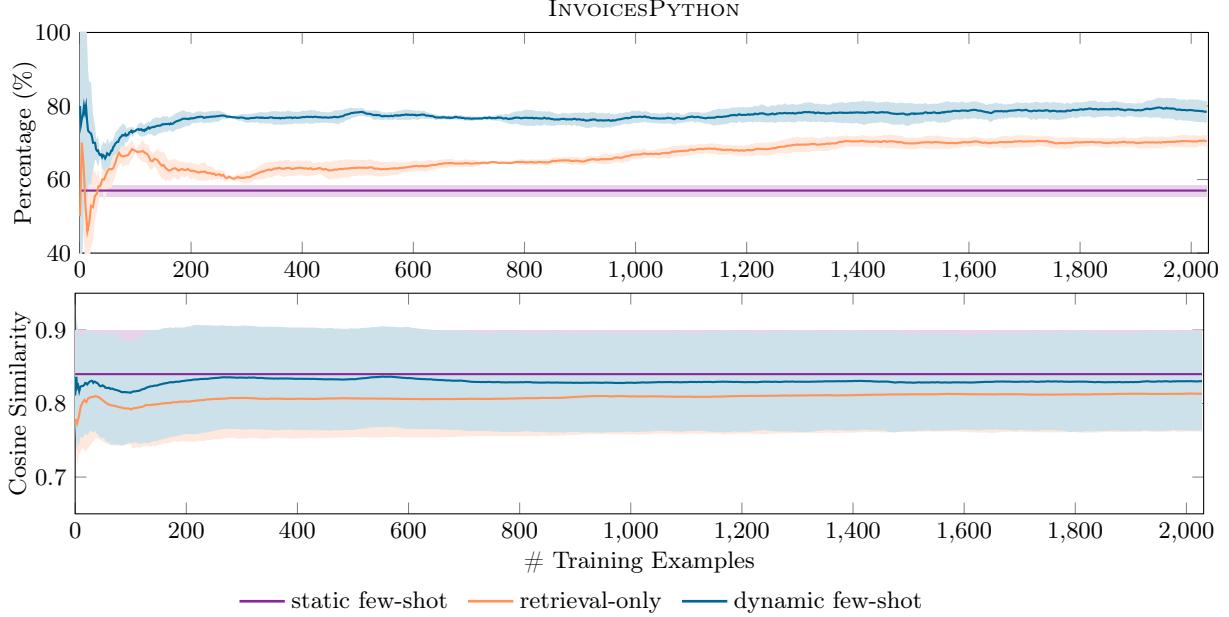


Figure 7: Percentage of answerable query suggestions and cosine similarity for a increasing number of labeled training examples on INVOICESPYTHON. We use a moving average over the last 50 queries and report mean and standard deviation.

Training Set Size and Self-Learning Figure 7 shows the suggestion performance for a varying number of training examples on the INVOICESPYTHON instance. Here, we choose a random order of the user queries. We generate a suggested query for every user query and evaluate its answerability and similarity to the original user query. We also self-learn on all previous user queries. After an initial period with high variance in similarity and answerability, dynamic few-shot learning quickly improves in answerability, suggesting that only 500 user queries are sufficient to effectively discern the answerability for suggested queries on our datasets. The answerability of the retrieval-only approach also improves. This shows that the robust retrieval approach of Section 3.2 can combat hallucinated responses. The static few-shot approach offers very high similarity while suffering low answerability.

6 Conclusion

We introduce and address the novel task of query suggestion for agentic retrieval-augmented generation (RAG), focusing on enhancing user interaction when initial queries are unanswerable. Unlike traditional query suggestion, our setting must reason about multi-step workflows and system capabilities without direct access to the agent’s internal knowledge or tools. To tackle this, we developed a robust dynamic few-shot learning framework that retrieves relevant answerability-aware examples via templating and embedding-based retrieval.

Our approach requires no manual labels, enabling practical deployment through self-learning on past user queries. Empirical results across three real-world RAG agents show that our method significantly outperforms both static few-shot and retrieval-only baselines in terms of both semantic similarity and answerability.

More broadly, our framework suggests a scalable path toward more agent-aware prompting: teaching LLMs to reason about tool affordances and system limitations through example-driven conditioning. Beyond query suggestion, our techniques offer building blocks for adaptive interaction, agent debugging, and workflow inspection in complex agentic systems.

References

- Tassallah Abdullahi, Ritambhara Singh, and Carsten Eickhoff. Retrieval augmented zero-shot text classification. In *ICTIR*, pages 195–203. ACM, 2024.
- Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. In-context examples selection for machine translation. In *ACL (Findings)*, pages 8857–8873. Association for Computational Linguistics, 2023.
- Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. Convai3: Generating clarifying questions for open-domain dialogue systems (clariq). volume abs/2009.11352, 2020.
- Jinheon Baek, Nirupama Chandrasekaran, Silviu Cucerzan, Allen Herring, and Sujay Kumar Jauhar. Knowledge-augmented large language models for personalized contextual query suggestion. In *WWW*, pages 3355–3366. ACM, 2024.
- Ricardo A. Baeza-Yates, Carlos A. Hurtado, and Marcelo Mendoza. Query recommendation using query logs in search engines. In *EDBT Workshops*, volume 3268 of *Lecture Notes in Computer Science*, pages 588–596. Springer, 2004.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020.
- Huanhuan Cao, Dixin Jiang, Jian Pei, Qi He, Zhen Liao, Enhong Chen, and Hang Li. Context-aware query suggestion by mining click-through and session data. In *KDD*, pages 875–883. ACM, 2008.
- Claudio Carpineto and Giovanni Romano. A survey of automatic query expansion in information retrieval. volume 44, New York, NY, USA, jan 2012. Association for Computing Machinery. doi: 10.1145/2071389.2071390. URL <https://doi.org/10.1145/2071389.2071390>.
- Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. RQ-RAG: learning to refine queries for retrieval augmented generation. volume abs/2404.00610, 2024.
- Heng Ding, Shuo Zhang, Darío Garigliotti, and Krisztian Balog. Generating high-quality query suggestion candidates for task-based search. In *ECIR*, volume 10772 of *Lecture Notes in Computer Science*, pages 625–631. Springer, 2018.
- Yi Dong, Ronghui Mu, Gaojie Jin, Yi Qi, Jinwei Hu, Xingyu Zhao, Jie Meng, Wenjie Ruan, and Xiaowei Huang. Position: Building guardrails for large language models requires systematic design. In *ICML*. OpenReview.net, 2024.
- Matthew Freestone and Shubhra Kanti Karmaker Santu. Word embeddings revisited: Do llms offer something new? volume abs/2402.11094, 2024.
- Darío Garigliotti and Krisztian Balog. Generating query suggestions to support task-based search. In *SIGIR*, pages 1153–1156. ACM, 2017.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. volume abs/2311.05232, 2023.
- Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama guard: Llm-based input-output safeguard for human-ai conversations. volume abs/2312.06674, 2023.

- Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joonsuk Park, and Jaewoo Kang. Tree of clarifications: Answering ambiguous questions with retrieval-augmented large language models. In *EMNLP*, pages 996–1009. Association for Computational Linguistics, 2023.
- Suyoun Kim, Duc Le, Weiyi Zheng, Tarun Singh, Abhinav Arora, Xiaoyu Zhai, Christian Fuegen, Ozlem Kalinli, and Michael L. Seltzer. Evaluating user perception of speech recognition system quality with semantic distance metric. In *INTERSPEECH*, pages 3978–3982. ISCA, 2022.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. CLAM: selective clarification for ambiguous questions with large language models. volume abs/2212.07769, 2022.
- Patrick S. H. 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. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*, 2020.
- Moxin Li, Wenjie Wang, Fuli Feng, Fengbin Zhu, Qifan Wang, and Tat-Seng Chua. Think twice before assure: Confidence estimation for large language models through reflection on multiple answers. volume abs/2403.09972, 2024.
- Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. Query rewriting for retrieval-augmented large language models. volume abs/2305.14283, 2023.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *NeurIPS*, 2023.
- Nitika Mathur, Timothy Baldwin, and Trevor Cohn. Putting evaluation in context: Contextual embeddings improve machine translation evaluation. In *ACL (1)*, pages 2799–2808. Association for Computational Linguistics, 2019.
- Qiaozhu Mei, Dengyong Zhou, and Kenneth Church. Query suggestion using hitting time. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM ’08, page 469–478, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781595939913. doi: 10.1145/1458082.1458145. URL <https://doi.org/10.1145/1458082.1458145>.
- Agnès Mustar, Sylvain Lamprier, and Benjamin Piwowarski. On the study of transformers for query suggestion. volume 40, pages 18:1–18:27, 2022.
- Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. 2023.
- Jessie Ooi, Xiuqin Ma, Hongwu Qin, and Siau Chuin Liew. A survey of query expansion, query suggestion and query refinement techniques. In *2015 4th International Conference on Software Engineering and Computer Systems (ICSECS)*, pages 112–117, 2015. doi: 10.1109/ICSECS.2015.7333094.
- OpenAI. GPT-4 technical report. volume abs/2303.08774, 2023.
- OpenAI. Embeddings, 2024. URL <https://platform.openai.com/docs/guides/embeddings>.
- Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive apis. volume abs/2305.15334, 2023.
- Shishir G. Patil, Tianjun Zhang, Vivian Fang, Noppapon C., Roy Huang, Aaron Hao, Martin Casado, Joseph E. Gonzalez, Raluca Ada Popa, and Ion Stoica. Goex: Perspectives and designs towards a runtime for autonomous LLM applications. volume abs/2404.06921, 2024.
- Shreya Rajpal. Guardrails ai, 2023. URL <https://www.guardrailsai.com/>.

- Traian Rebedea, Razvan Dinu, Makesh Narsimhan Sreedhar, Christopher Parisien, and Jonathan Cohen. Nemo guardrails: A toolkit for controllable and safe LLM applications with programmable rails. In *EMNLP (Demos)*, pages 431–445. Association for Computational Linguistics, 2023.
- Matthew Renze and Erhan Guven. Self-reflection in llm agents: Effects on problem-solving performance, 2024.
- Antonio Rodriguez, Anubhav Mishra, and Dani Mitchell. Safeguard a generative ai travel agent with prompt engineering and amazon bedrock guardrails, 2024. URL <https://aws.amazon.com/blogs/machine-learning/safeguard-a-generative-ai-travel-agent-with-prompt-engineering-and-amazon-bedrock-guardrails/>. Accessed on 18 June 2024.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *NeurIPS*, 2023.
- Yisheng Song, Ting Wang, Puyu Cai, Subrota K. Mondal, and Jyoti Prakash Sahoo. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. volume 55, pages 271:1–271:40, 2023.
- Michael Tänzer, Sebastian Ruder, and Marek Rei. Memorisation versus generalisation in pre-trained language models. In *ACL (1)*, pages 7564–7578. Association for Computational Linguistics, 2022.
- Anuja Tayal and Aman Tyagi. Dynamic contexts for generating suggestion questions in RAG based conversational systems. In *WWW (Companion Volume)*, pages 1338–1341. ACM, 2024.
- Laurens van der Maaten. Accelerating t-sne using tree-based algorithms. volume 15, pages 3221–3245, 2014.
- Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karanikolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. Supernaturalinstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *EMNLP*, pages 5085–5109. Association for Computational Linguistics, 2022.
- Hamed Zamani, Susan T. Dumais, Nick Craswell, Paul N. Bennett, and Gord Lueck. Generating clarifying questions for information retrieval. In *WWW*, pages 418–428. ACM / IW3C2, 2020.
- Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. RAFT: adapting language model to domain specific RAG. volume abs/2403.10131, 2024.

A Workflow Embeddings

Figure 8 shows the “workflow embeddings,” i.e. the embeddings that we obtained from templating a set of real-world user queries as described in Section 3.1. The query answerability is decided by the LLM itself, which makes up the self-learning component of our approach. We can discern groups of answerable and non-answerable questions. The zoomed-in region focuses on user queries about late payments. Without a Python tool, ad-hoc calculations such as averages cannot be evaluated. This shows that discerning the answerability of a query requires a nuanced understanding of the RAG, but this can be delivered to the LLM via dynamic few-shot examples through our retrieval approach as described in Section 3.2.

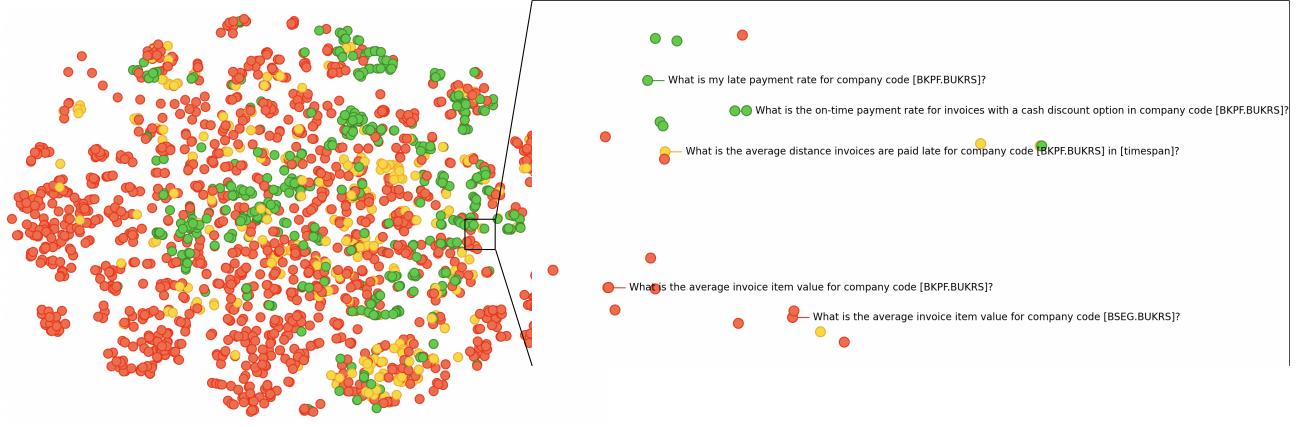


Figure 8: Workflow embedding: TSNE (van der Maaten, 2014) of template embeddings for 2029 user questions of the INVOICESNoPYTHON dataset. Red, yellow, and green dots indicate queries that have no workflow, no knowledge, and are answerable, respectively.

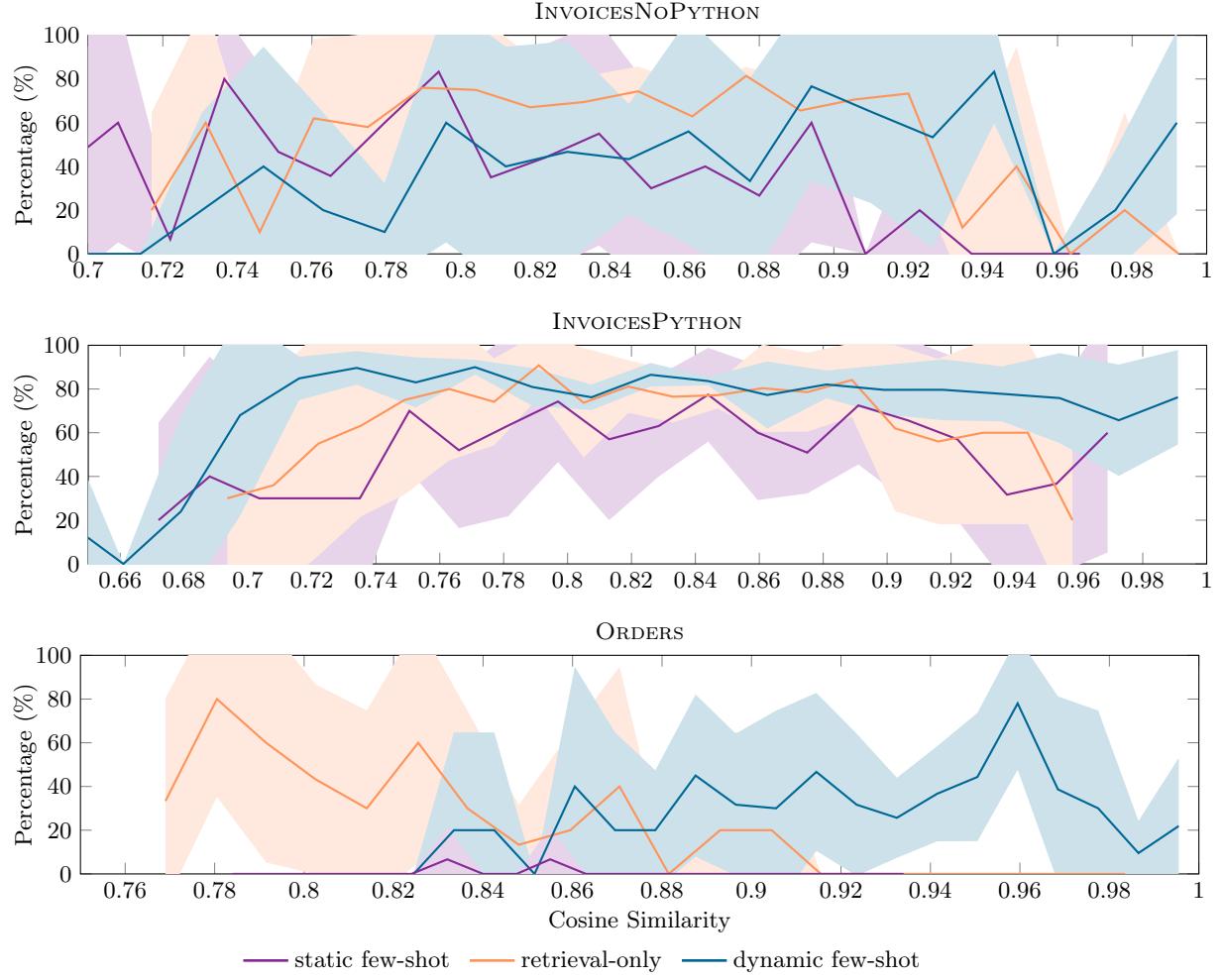


Figure 9: Similarity and answerability of the suggested queries. We show the percentage of queries that are classified as answerable, grouped by their cosine similarity to the original query. We report mean and standard deviation.