

# CHATTY-KG: A Multi-Agent AI System for On-Demand Conversational Question Answering over Knowledge Graphs

Reham Omar  
Concordia University  
Canada

Abdelghny Orogat  
Concordia University  
Canada

Ibrahim Abdelaziz  
IBM Research  
USA

Omij Mangukiya  
Concordia University  
Canada

Panos Kalnis  
KAUST  
Saudi Arabia

Essam Mansour  
Concordia University  
Canada

## Abstract

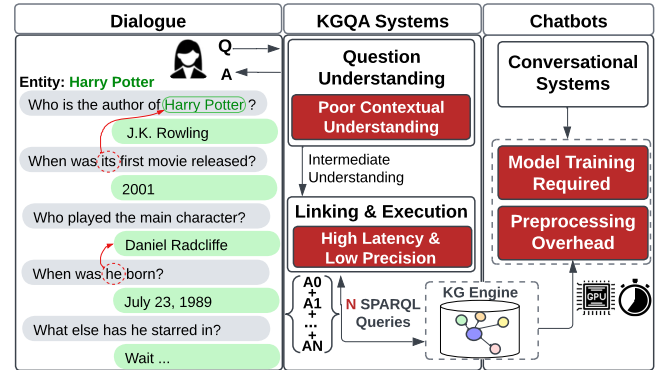
Conversational Question Answering over Knowledge Graphs (KGs) combines the factual grounding of KG-based QA with the interactive nature of dialogue systems. KGs are widely used in enterprise and domain applications to provide structured, evolving, and reliable knowledge. Large language models (LLMs) enable natural and context-aware conversations, but lack direct access to private and dynamic KGs. Retrieval-augmented generation (RAG) systems can retrieve graph content but often serialize structure, struggle with multi-turn context, and require heavy indexing. Traditional KGQA systems preserve structure but typically support only single-turn QA, incur high latency, and struggle with coreference and context tracking. To address these limitations, we propose CHATTY-KG, a modular multi-agent system for conversational QA over KGs. CHATTY-KG combines RAG-style retrieval with structured execution by generating SPARQL queries through task-specialized LLM agents. These agents collaborate for contextual interpretation, dialogue tracking, entity and relation linking, and efficient query planning, enabling accurate and low-latency translation of natural questions into executable queries. Experiments on large and diverse KGs show that CHATTY-KG significantly outperforms state-of-the-art baselines in both single-turn and multi-turn settings, achieving higher F1 and P@1 scores. Its modular design preserves dialogue coherence and supports evolving KGs without fine-tuning or pre-processing. Evaluations with commercial (e.g., GPT-4o, Gemini-2.0) and open-weight (e.g., Phi-4, Gemma 3) LLMs confirm broad compatibility and stable performance. Overall, CHATTY-KG unifies conversational flexibility with structured KG grounding, offering a scalable and extensible approach for reliable multi-turn KGQA.

## ACM Reference Format:

Reham Omar, Abdelghny Orogat, Ibrahim Abdelaziz, Omij Mangukiya, Panos Kalnis, and Essam Mansour. 2025. CHATTY-KG: A Multi-Agent AI System for On-Demand Conversational Question Answering over Knowledge Graphs. In . ACM, New York, NY, USA, 15 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

## 1 Introduction

Conversational Question Answering over Knowledge Graphs (KGs) aims to combine the factual accuracy and structured reasoning of KG-based question answering with the natural, user-friendly



**Figure 1: Limitations of current KGQA and chatbot-based systems for real-time conversational access to arbitrary KGs. KGQA systems face issues with contextual understanding, query fragmentation, and latency. Chatbots offer better conversational QA over KGs but require expensive training and preprocessing, limiting adaptability and scalability.**

experience of dialogue systems. KGs are widely adopted in domain-specific and enterprise applications to integrate heterogeneous data sources and provide reliable and up-to-date knowledge. Examples include general KGs, such as Wikidata <sup>1</sup>, biomedical graphs like BioRDF <sup>2</sup> and UMLS <sup>3</sup>, academic KGs, such as DBLP <sup>4</sup>, and legal or financial graphs like FinDKG <sup>5</sup>. Despite their strengths, interacting with KGs typically requires formal query languages, such as SPARQL or Cypher, creating a steep learning curve for non-expert users. This limits the deployment of conversational agents that can reliably answer questions over these knowledge sources.

To make KGs more accessible, Knowledge Graph Question Answering (KGQA) systems enable users to ask questions in natural language. However, most systems rely on domain-specific pipelines, require heavy KG preprocessing (e.g., EDGQA [15]), and depend on finetuned models (e.g., KGQAn [30]), which limits their scalability and adaptability. These systems are typically designed for single-turn questions and often suffer from high latency. Such limitations make them unsuitable for multi-turn conversations that require context tracking and efficient processing. As shown in Figure 1, they also struggle with contextual understanding and multi-query

<sup>1</sup><https://www.wikidata.org/>

<sup>2</sup><https://bio2rdf.org/>

<sup>3</sup><https://www.nlm.nih.gov/research/umls>

<sup>4</sup><https://dblp.org/>

<sup>5</sup><https://xiaohui-victor-li.github.io/FinDKG/>

coordination, resulting in low accuracy and slow response times, as demonstrated by KGQAn [30] and EDGQA [15].

Recent advancements in large language models (LLMs), such as GPT-4 [32] and Gemini [42], have greatly improved natural language understanding and reasoning. One might expect retrieval-augmented generation (RAG) to remove the need for explicit KG access by retrieving a subgraph and using it as context for the LLM. However, there is a gap between the RAG *paradigm* and its current *implementations*. In practice, graph-RAG systems such as Microsoft GraphRAG [13] must first interpret the question and extract entities before retrieval, but retrieval itself depends on heavy graph indexing and embedding stores that consume large memory and do not scale to large KGs. The retrieved subgraphs are then serialized into text, which discards structural information and weakens answer precision, especially for multi-entity or list-style queries. These systems may lack dialogue support, preventing multi-turn refinement or context reuse. As illustrated in Figure 1, chatbot-based KG access (e.g., CONVINSE [5], EXPLAIGNN [6]) requires retraining or KG-specific preprocessing to remain accurate, making it costly and inflexible for evolving graphs. These limitations show that RAG-as-a-paradigm is promising, but current graph-RAG implementations fail to deliver accurate, structure-aware, and dialogue-capable KG access.

To address these limitations, we introduce CHATTY-KG, a modular multi-agent framework for conversational KGQA. The key technical novelty lies in combining RAG-as-a-paradigm with structured query execution through a training-free multi-agent design that requires no KG-specific preprocessing and generalizes across arbitrary graphs. Instead of asking an LLM to infer an answer from serialized triples, CHATTY-KG uses LLM agents to generate SPARQL queries over the KG, preserving graph semantics and reducing hallucinations. The pipeline is decomposed into lightweight agents for question understanding, dialogue context tracking, entity linking, and query generation, coordinated through a controller for low-latency execution. CHATTY-KG leverages recent LLM advances together with orchestration frameworks like LangGraph<sup>6</sup> while avoiding costly retraining. This design enables real-time, verifiable access to *evolving KGs* and supports both single-turn and multi-turn dialogues with minimal overhead.

Our extensive evaluation confirms that CHATTY-KG achieves outstanding performance across five real-world and large KGs of diverse application domains. It consistently outperforms state-of-the-art KGQA systems, including KGQAn [30] for single-turn questions, RAG-based systems, such as GraphRAG [13] and ColBERT [21, 37], and chatbot systems, such as GPT-4o, Gemini-2.0, CONVINSE [5] and EXPLAIGNN [6] for multi-turn dialogue. CHATTY-KG's modular design supports contextual reasoning and coherent dialogue without domain tuning. Experiments across a wide range of commercial (e.g., GPT-4o, Gemini-2.0) and open-weight LLMs (e.g., Phi-4, Gemma-3) demonstrate strong compatibility and stable performance. In general, CHATTY-KG offers a scalable, extensible, and reliable framework for conversational KGQA in real time with significant improvements in accuracy, efficiency, and adaptability.

In summary, our contributions are:

- A modular multi-agent architecture for conversational KGQA, where individual agents can be swapped or enhanced independently, supporting future extensibility.
- Supports single- and multi-turn conversations with LLM-powered context tracking and disambiguation, while maintaining low latency for interactive use.
- Efficient SPARQL-based KG access via a query planning agent that eliminates the need for offline preprocessing.
- A comprehensive evaluation across five real KGs from diverse domains demonstrates CHATTY-KG's superior performance in answer correctness and response time compared to state-of-the-art systems, with consistent results across commercial and open-weight LLMs.

## 2 Background and Related Work

This section surveys related work on KGQA, conversational models, and LLM-based agent frameworks. We outline the challenges of integrating conversational capabilities with structured knowledge and motivate our training-free and LLM-driven approach.

### 2.1 Knowledge Graph Question Answering

Conversational question answering over KGs typically requires three main steps: (i) Dialogue Understanding, which creates an abstract representation from the key entities and relations extracted from the question and conversation history; (ii) Linking, which maps the abstract representation to vertices and predicates from the KG; and (iii) Answer Generation, which uses the abstract representation and the mapped KG vertices and predicates to either extract the answer from the KG or generate it using a trained model. In this section, we discuss the two categories of systems summarized in Table 1: KGQA; e.g., [15, 30], and Conversational Systems; e.g., [5, 6] and how it compares to our approach.

**KGQA systems** rely on generating SPARQL queries that represent an individual question. Examples include gAnswer [14, 54], NSQA [20], and WDAqua [10]. The top two performing systems, KGQAn [30] and EDGQA [15], are examined in greater detail in this section to better understand their design choices, strengths, and limitations. KGQAn is the state-of-the-art across multiple datasets [30], including QALD-9 [44], while EDGQA achieves strong results on the LC-QuAD 1.0 dataset [15, 43]. However, these systems cannot handle conversational interactions as they are designed for single-turn and standalone questions. KGQAn's pretrained models are limited by their training on single-turn queries and a relatively small context window length that cannot fit conversation history. Similarly, EDGQA's rule-based approach does not scale to multi-turn conversations. Both KGQAn and EDGQA have high response times, making them unfit for conversational use. KGQAn's latency is due to repeated similarity-server requests, numerous SPARQL queries execution for high recall, and an answer-type filtering step. EDGQA also runs multiple SPARQL queries and uses three different systems for linking, contributing to its delay [30].

For question understanding, KGQAn fine-tunes an LLM to generate triples from questions, requiring a manually created dataset that is time-consuming and subject to annotator bias. It also trains a separate answer type detection model. In contrast, EDGQA uses the Stanford CoreNLP parser with rules tailored to a specific dataset,

<sup>6</sup><https://python.langchain.com/docs/langgraph/>

**Table 1: Comparison of existing KGQA and conversational QA systems across key design aspects. KGQA systems lack multi-turn dialogue support and suffer from high latency, while conversational systems depend on fine-tuning and KG-specific preprocessing. CHATTY-KG overcomes these limitations with a LLM-powered multi-agent approach for real-time performance.**

System	Dialogue	Response Time	Question Understanding	Linking	Training	Pre-processing
KGQAn[30]	x	high	Fine-tuning	Semantic similarity based	✓	x
EDGQA[15]	x	high	Constituency parsing	Index-based	x	✓
CONVINSE[5]	✓	low	Fine-tuning	Index-based	✓	✓
EXPLAIGNN[6]	✓	low	Fine-tuning	Index-based	✓	✓
CHATTY-KG	✓	low	LLM-Powered	LLM-Powered	x	x

limiting cross-domain applicability. For linking, KGQAn makes hundreds of similarity-server calls, slowing response time, while EDGQA uses three indexing systems (Falcon [36], EARL [12], Dexter [3]), requiring heavy pre-processing and reducing adaptability to new knowledge graphs. Overall, KGQAn demands extensive model training, while EDGQA relies on significant pre-processing.

**Conversational systems** rely on information retrieval techniques to extract or generate answers. CONVINSE [5] and EXPLAIGNN [6] are end-to-end conversational question-answering systems on heterogeneous data sources, including KGs, text, and tables. These systems, which only support the Wikidata KG [46], handle conversational interactions with low response times, making them suitable for real-time user interaction. For question understanding, they transform questions into intent-explicit structured representations (SR) using fine-tuned transformer models [45] (BART [25] and T5 [33]). CONVINSE and EXPLAIGNN formalize linking as evidence retrieval and scoring. They use Clocq [4] for named entity disambiguation and KG-item retrieval. Clocq requires KG pre-processing before KG facts retrieval. For answer generation, CONVINSE trains a Fusion-in-Decoder (FiD) model [17] to generate answers based on top-ranked evidence. EXPLAIGNN constructs a heterogeneous answering graph from retrieved entities and evidence. Both systems require model training for question understanding and answer generation. They also require KG pre-processing, which limits their adaptability to new KGs.

In summary, existing systems face key limitations that hinder effective conversational question answering over diverse and evolving KGs. Traditional KGQA systems lack support for multi-turn dialogue and suffer from high latency. Conversely, conversational systems improve interactivity and speed but depend heavily on fine-tuning and KG-specific pre-processing, reducing flexibility across domains. In contrast, CHATTY-KG introduces a modular, low-latency conversational framework that eliminates the need for training or pre-processing. It leverages LLMs for dynamic dialogue tracking, question understanding, and entity linking to enable scalable and adaptable QA over arbitrary knowledge graphs.

## 2.2 LLM Prompting Strategies

LLMs such as GPT-4, Gemini, and Qwen [51] have shown strong language understanding and generation skills across diverse tasks. Their generalization without task-specific training makes them especially valuable in modular systems like CHATTY-KG, where they serve as interchangeable agents for reasoning, classification, or generation. Prompt engineering has emerged as a lightweight, cost-effective alternative to fine-tuning by steering LLMs through

carefully crafted instructions [40]. Common prompting strategies include: 1) zero-shot prompting with only a task description is provided to the LLM, 2) few-shot prompting which adds to the LLM a handful of examples, and 3) chain-of-thought (CoT) prompting, which guides intermediate reasoning steps before producing an answer [22, 49]. The best strategy depends on the task: classification or retrieval often works well with zero-shot prompts, while generation or complex reasoning benefits from few-shot or CoT prompting. In CHATTY-KG, simpler agents (e.g., for classification or routing) use zero-shot prompts for speed, while agents handling contextual understanding or answer synthesis apply few-shot or CoT prompting to support multi-step reasoning. This flexibility optimizes both efficiency and quality across the multi-agent pipeline.

## 2.3 LLM Agents

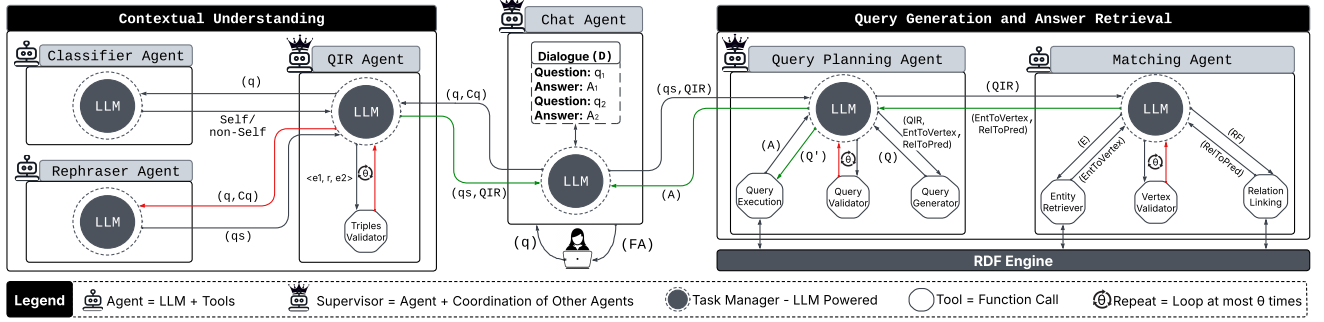
LLM agents are emerging systems that use the reasoning, planning, and language capabilities of large language models to autonomously complete tasks. At their core, these agents treat the LLM as a central decision-maker that interacts with tools, APIs, and environments in a loop of observation, reasoning, and action. Recent frameworks, such as ReAct [53], AutoGPT [39], and BabyAGI [29], combine LLMs with tool use and planning mechanisms. More structured frameworks like LangChain, LlamaIndex<sup>7</sup>, and OpenAgents [50] provide abstractions for integrating tools, retrievers, memory, and control flows. LangChain enables modular pipelines that compose LLMs with APIs, databases, and reasoning components. LangGraph<sup>8</sup>, a recent LangChain extension, introduces stateful graph-based control for defining multi-agent workflows with dynamic routing.

A key enabler of these systems is tool calling (also known as function calling), where the LLM invokes external tools at inference time to offload specific tasks, such as computation, API queries, or structured data access. Tool calling has become central to modern agent frameworks, allowing models to extend beyond generation and act as orchestrators. In our system, agents selectively call tools based on their roles, for example, to resolve entity mentions, fetch KG sub-graphs, or execute SPARQL queries. This targeted, task-specific integration enables agents to remain lightweight and model-agnostic, while grounding their decisions in accurate, up-to-date KG content.

LLM agents are increasingly used in domains, such as web automation [52], code generation [48], and data wrangling [26]. In these settings, they break down complex goals, retrieve intermediate results, and adapt to user feedback [47]. However, most systems are tailored to unstructured or semi-structured data and struggle

<sup>7</sup><https://www.llamaindex.ai/>

<sup>8</sup><https://python.langchain.com/docs/langgraph/>



**Figure 2: CHATTY-KG’s hierarchical multi-agent architecture. A top-level Chat Agent coordinates two supervised modules: Contextual Understanding and Query Generation & Answer Retrieval, each composed of specialized LLM-powered agents. This design enables modular, low-latency KGQA without training or pre-processing, and improves adaptability and real-time performance.**

with formal sources, such as relational databases or knowledge graphs (KGs) [38]. Challenges in schema alignment, entity linking, and precise query generation remain underexplored. Therefore, the use of LLM agents for KG question answering (KGQA) is still emerging. In this work, we demonstrate how core agent capabilities, such as contextual understanding, linking, and modular reasoning, can be applied to KGQA in a lightweight and efficient manner. Our system avoids complex planning or orchestration. Instead, our agents use prompt-guided LLM calls to interpret questions, link KG elements, and generate a minimal set of executable queries.

### 3 CHATTY-KG Architecture

CHATTY-KG adopts a hierarchical design that splits the task into two main subtasks: language-level processing and graph-level reasoning. Language-level tasks include question classification, ambiguity resolution, and intermediate representation generation. Graph-level tasks cover entity/relation linking and SPARQL query construction. Each subtask is handled by a cluster of LLM agents. Each agent wraps a *task-managed LLM*, a language model guided by role-specific logic and tools. Some agents act as supervisors, delegating to subordinate agents. This modular and domain-agnostic architecture supports accurate reasoning, flexible operation across KGs, and efficient execution without retraining or KG-specific preprocessing. The matching agent does not assume a fixed schema and instead dynamically extracts relevant subgraphs using SPARQL queries. Key stages incorporate validation and retry mechanisms to mitigate error propagation. We first define the main representations.

**Definition 3.1 (Dialogue).** The dialogue history is defined as  $\mathcal{D} = \{(q_1, A_1), (q_2, A_2), \dots, (q_{n-1}, A_{n-1})\}$ , where each  $q_i$  is a user-issued question and  $A_i = [a_1, a_2, \dots, a_m]$  is the corresponding system-generated answer. Each  $a_j$  may be a count, a Boolean value, or an entity;  $A_i$  may also be a list of such values.

**Definition 3.2 (Question Intermediate Representation (QIR)).** Given a question  $q$  that mentions entities, expresses relations, and targets an unknown variable, potentially involving intermediate variables for multi-hop reasoning, the QIR is defined as  $QIR = (E, U, R, RF)$ . Here,  $E$  is the set of mentioned entities,  $U$  is the set of variables (including the target and any intermediates),  $R$  is the set of relation phrases in  $q$ , and  $RF$  is a set of relational facts. Each  $rf \in RF$  is a triple  $\langle e_1, r, e_2 \rangle$ , where  $e_1, e_2 \in (E \cup U)$  and  $r \in R$ .

CHATTY-KG is structured into three core modules: *Contextual Understanding*, *Chat Management*, and *Query Generation and Answer Retrieval* (Figure 2). Each module contains one or more agents and tools with well-defined communication interfaces. The system operates on a dialogue  $\mathcal{D}$ , using its history to form the context  $C_q$  for interpreting new user questions. Upon receiving a new question  $q_i$ , the **Chat Agent** initiates processing by forwarding  $q_i$  and  $C_q$  to the **QIR Agent**, which builds the intermediate representation to guide query generation. To manage complexity, the QIR Agent delegates two subtasks to specialized agents. First, the **Classifier Agent** determines whether  $q_i$  is *self-contained* (interpretable alone) or *context-dependent* (requiring prior turns). For example, “Who is the author of Harry Potter?” is self-contained, while “When was its first movie released?” depends on earlier context.

The classification ensures downstream agents receive clear and unambiguous input. If the question is classified as context-dependent, the QIR Agent forwards both  $q_i$  and  $C_q$  to the **Rephraser Agent**. This agent reformulates the question into a self-contained version using dialogue context. For example, it transforms “When was its first movie released?” into “When was the first Harry Potter movie released?”. The QIR Agent then processes the resulting question, whether rephrased or original, to generate the Question Intermediate Representation (QIR). This involves extracting a set of triples representing the question’s semantics.

CHATTY-KG uses an assertion-based validation approach to ensure output correctness. Inspired by software assertions, it defines task-specific structural and semantic checks that are independent of input content. For example, the **Triples Validator** checks that each extracted triple has valid components (subject, predicate, object). If validation fails, the QIR Agent retries generation up to a threshold  $\theta$ . Once validated, the triples form the QIR. For instance, the QIR for the rephrased question is  $\{E: \{“Harry Potter”\}, U: \{“?year”\}, RF: \{\langle Harry Potter, released, ?year \rangle\}\}$ .

The **Query Planning Agent** receives the QIR and identifies a minimal set of SPARQL queries to retrieve the answer. It begins by delegating entity and relation linking to the **Matching Agent**, which uses the **Entity Retriever** and **Relation Linking** tools to map QIR elements to KG URIs. The resulting mappings are passed to the **Vertex Validator**, which checks that each entity is syntactically correct and appears in the candidate list. If validation fails, the system retries up to  $\theta$  times. After successful validation, the URIs are returned to the planning agent for query construction.

The **Query Generator** constructs candidate SPARQL queries. The **Query Planning Agent** selects a minimal subset for execution, typically one to three queries, based on its understanding, without relying on a fixed threshold. The selected queries are passed to the **Query Validator**, which uses assertions to ensure syntactic and semantic correctness, retrying on failure. This avoids executing irrelevant queries and reduces overhead. The selected queries are sent to the **Query Execution** module to retrieve the raw answer  $A$ . The Chat Agent may rephrase it into a user-friendly response  $FA$ , e.g., turning "2001" into "The first Harry Potter movie was released in 2001." The final answer is returned to the user, and the dialogue  $\mathcal{D}$  is updated to complete the interaction.

## 4 Chat Agent and Contextual Understanding

This section introduces two core components: the *Chat Agent*, which manages dialogue state and controls workflow, and the *Contextual Understanding* module, which transforms natural language questions into a structured Question Intermediate Representation.

### 4.1 Chat Agent

The **Chat Agent** acts as the top-level controller in CHATTY-KG's multi-agent framework. It is responsible for: (i) orchestrating module execution, (ii) maintaining dialogue history, and (iii) reformulating answers and enabling multilingual queries through LLM-based translation. When a new user question  $q_n$  arrives, the Chat Agent retrieves the dialogue history  $\mathcal{D}$  and constructs the corresponding context  $C_q$  needed to interpret the question. Since  $C_q$  is passed as input to prompt-based LLM agents (e.g., the Rephraser Agent), it must remain within the LLM's token limit. We define  $C_q = \{(q_i, A_i^L)\}_{i=1}^{n-1}$ , where each  $A_i^L$  is a truncated prefix of  $A_i$ , limited to  $L$  items when  $|A_i| > L$ . This preserves essential context while ensuring compact prompts. The pair  $(q_n, C_q)$  is passed to the Contextual Understanding module, which generates the structured Question Intermediate Representation (QIR). The QIR is then forwarded to the Query Generation and Answer Retrieval module, which constructs and executes the corresponding SPARQL query and returns the raw answer  $A_n$ . Optionally, the Chat Agent reformulates  $A_n$  into a more natural final answer  $FA_n$  using a zero-shot prompting strategy:

$$FA_n = f(A_n, q_n, I_{FA}), \quad (1)$$

where  $I_{FA}$ <sup>9</sup> is a prompt instructing the LLM to express short literal answers in fluent natural language. We adopt zero-shot prompting due to the simplicity and consistency of the task, which typically involves converting short outputs into complete sentences. Finally, the Chat Agent appends  $(q_n, A_n)$  to the dialogue history and awaits the next question. This process enables context-aware, multi-turn interaction while maintaining low latency and prompt efficiency.

### 4.2 Contextual Understanding

To interpret a user question  $q_i$  in its dialogue context  $C_q$ , the Contextual Understanding module runs a multi-stage pipeline that converts natural language into a structured Question Intermediate Representation (QIR). This process is managed by the *QIR Agent*, which coordinates two key subtasks handled by auxiliary agents:

---

#### Algorithm 1 Contextual Understanding Pipeline

---

**Input:**  $q_i$ : User question,  $C_q$ : Question context,  $\theta$ : Retry threshold,  $I_C$ : Classification instructions,  $I_{RF}$ : Reformulation instructions,  $I_{QU}$ : Understanding instructions

**Output:**  $QIR = (E, U, R, RF)$ : Intermediate Representation

```

1:  $q_{type} \leftarrow \text{ClassifierAgent}(q_i, I_C)$ 
2: if  $q_{type} = \text{'Dependent'}$  then
3:    $q_{is} \leftarrow \text{RephraserAgent}(q_i, C_q, I_{RF})$ 
4: else
5:    $q_{is} \leftarrow q_i$ 
6: end if
7:  $valid \leftarrow \text{False}$ 
8: while  $valid = \text{False} \wedge \theta > 0$  do
9:    $triples \leftarrow \text{QIRAgent.GenerateTriples}(q_{is}, I_{QU})$ 
10:   $valid \leftarrow \text{QIRAgent.ValidateTriples}(triples)$ 
11:   $\theta \leftarrow \theta - 1$ 
12: end while
13:  $QIR \leftarrow \text{QIRAgent.ConstructQIR}(triples)$ 
14: return  $QIR$ 

```

---

classification and reformulation. As shown in Algorithm 1, the *Classifier Agent* first determines whether  $q_i$  is self-contained or context-dependent. If the question depends on prior context, the *Rephraser Agent* generates a standalone version  $q_{is}$  using  $C_q$  to resolve references and omissions. Otherwise,  $q_{is} \leftarrow q_i$ .

The unified question  $q_{is}$  is then processed by the QIR Agent to extract semantic triples via prompt-guided LLM interaction. These triples are validated for structure and meaning, and the agent retries up to  $\theta$  times if validation fails. Once validated, the resulting QIR is returned to the Chat Agent for downstream processing. Separating these subtasks improves LLM performance by reducing prompt complexity and avoiding instruction overload, as each agent uses a dedicated session and objective-specific prompt.

**4.2.1 Classifier Agent.** The Classifier Agent identifies whether a question is *self-contained* or *context-dependent* by detecting unresolved references or pronouns that imply reliance on earlier dialogue turns. It uses a zero-shot prompting strategy with a natural-language instruction  $I_C$  that explains the classification criteria and guides the decision. We avoid using traditional machine learning classifiers, which require labeled training data that is costly and often unavailable. Moreover, this task goes beyond feature-level classification such as pronoun presence. For instance, "When was the first movie released?" appears complete but is context-dependent if the subject entity was mentioned earlier. The LLM-based approach enables semantic reasoning over such implicit dependencies. The agent returns a label  $q_{type} \in \{\text{Self-contained}, \text{Dependent}\}$ , which determines the subsequent processing path. The classification step is formalized as:

$$q_{type} = f(q_i, I_C) \quad (2)$$

**4.2.2 Rephraser Agent.** The Rephraser Agent converts a context-dependent question  $q_i$  into a self-contained version  $q_{is}$  by resolving implicit references using the dialogue history  $C_q$ . This is achieved through a prompt-driven LLM function:

<sup>9</sup>All prompts are available in the [supplementary materials](#).



$$q_{is} = f(q_i, C_q, I_{RF}) \quad (3)$$

Here, the model receives the question  $q_i$ , its context  $C_q$ , and an instruction prompt  $I_{RF}$  that specifies the criteria for semantic completeness. The prompt directs the model to resolve dependencies—such as pronouns, ellipses, or anaphora—and rewrite the question as a standalone utterance that no longer depends on prior turns. Unlike rule-based systems that rely on handcrafted heuristics, this approach leverages the LLM’s pretrained reasoning to generalize across varied linguistic forms. Although the task involves text generation, it is bounded in scope: the objective is to restructure the input without introducing new content. This constraint makes it well-suited for zero-shot prompting. The resulting question  $q_{is}$  preserves the original intent of  $q_i$  while satisfying structural and semantic requirements for QIR generation.

Ambiguous or conflicting entity mentions across dialogue turns are resolved by the rephraser agent using the full dialogue history. The history stores explicit entity mentions in  $C_q$ , allowing the agent to replace pronouns with the correct referenced entity before reformulating the question. Since the agent relies on LLMs with strong long-context coreference abilities, it can disambiguate pronouns and produce a clear, standalone version of the user query.

**4.2.3 QIR Agent.** The QIR Agent converts the self-contained question  $q_{is}$  into a structured semantic form. It generates a set of triples  $RF = \{\langle s_i, r_i, o_i \rangle\}_{i=1}^k$ , where  $s_i, o_i \in E \cup U$  and each relation  $r_i \in \mathcal{R}$ , the set of relation phrases in  $q_{is}$ . These triples represent the core semantic structure of the question and are incorporated into the final QIR only after passing a validation step. An initial prompting method guided the LLM to extract triples by clarifying variable-entity distinctions, highlighting relevant facts, and enforcing output format. However, this led to frequent issues such as omitted variables and generic or ambiguous predicates (e.g., *has*) that distorted intent. To address this, the QIR Agent adopts a chain-of-thought (CoT) prompting strategy combined with few-shot learning. The CoT prompt decomposes the task into three reasoning steps: (i) identifying key entities, (ii) detecting unknowns (answers or intermediate variables), and (iii) generating verb- or noun-based relations between them. This structured reasoning improves precision and mitigates earlier shortcomings. Two examples (one with verb-based, one with noun-based relations) are included to guide generation. The agent receives a declarative prompt  $I_{QU}$  and example set  $e$ , returning a structured output in JSON format for downstream use. The process is formalized as:

$$RF = f(q_{is}, I_{QU}, e) \quad (4)$$

**Triples Validator.** After generating the candidate triples, the Triples Validator checks the output against structural and semantic constraints. Specifically, the result must satisfy: (1) it is a well-formed JSON object, (2) each triple includes a valid subject, predicate, and object, (3) at least one triple references a known entity, and (4) for non-boolean questions, at least one variable must appear. Boolean questions such as “Is Michelle the wife of Barack Obama?” are exempt from the fourth condition, as they may consist entirely of grounded triples like  $\langle \text{Michelle}, \text{wife of}, \text{Barack Obama} \rangle$ . If the output fails validation, the QIR Agent retries generation up to a

threshold  $\theta$ . Once validated, the QIR is constructed from the validated triples and returned.

**4.2.4 Time and Space Complexity.** Algorithm 1 executes a bounded sequence of LLM agent calls, whose cost is dominated by language model interactions. Let  $C_{LLM}$  denote the cost of one LLM invocation. The algorithm performs at most three LLM calls: one to the Classifier Agent (Line 1), one conditional call to the Rephraser (Line 3), and up to  $\theta$  calls to the QIR Agent for triple generation (Line 9). Thus, the total time complexity is  $O(\theta \cdot C_{LLM})$ , since the number of LLM calls is constant. The validation step (Line 10) uses lightweight rule-based checks and runs in constant time. Space complexity is dominated by the dialogue history  $\mathcal{D}$ , which adds one QA pair per turn, resulting in  $O(n)$  space for  $n$  turns.

## 5 Query Generation and Answer Retrieval

This section describes the *Query Planning Agent*, which converts the QIR into a small set of SPARQL queries to retrieve the answer  $A$ . Entity and relation linking are delegated to the *Matching Agent*. The agent executes only a high-confidence subset of candidate queries, requiring no KG-specific preprocessing, remaining compatible with standard RDF backends, and reducing latency.

### 5.1 Matching Agent

Given the  $QIR = (E, U, R, RF)$ , the Matching Agent aligns QIR entities and relations to their corresponding URIs in the target KG by querying its SPARQL endpoint at runtime. As shown in Algorithm 2, for each entity  $e \in E$  (Line 2), the *Entity Retriever* issues a keyword-based SPARQL query to retrieve up to  $v_{limit}$  candidate vertices (Line 3). Vertex selection is then performed via prompt-guided reasoning using instructions  $I_{VL}$  (Line 6). The result is validated by the *Vertex Validator*, which checks both membership in the candidate set and JSON correctness (Line 7). If validation fails, selection is retried up to  $\theta$  times (Lines 5–9) to mitigate LLM hallucinations or formatting errors. Validated mappings are stored in *EntToVertex* (Line 10). Next, for each relation  $r \in R$  (Line 12), the *Relation Linking Tool* retrieves and ranks candidate predicates to maximize recall. The resulting ranked list is stored in *RelToPred* (Line 14). Finally, the agent returns: (i) *EntToVertex* :  $E \rightarrow V$ , mapping each entity to a KG vertex, and (ii) *RelToPred* :  $R \rightarrow \text{List}(P)$ , mapping each relation to a list of semantically relevant predicates<sup>10</sup>. This runtime-only, modular design avoids KG-specific preprocessing, schema alignment, or offline indexing. As a result, it generalizes well across diverse KGs while maintaining high precision and adaptability.

**5.1.1 Entity Retriever.** Entity linking aims to match each known entity  $e \in E$  in the  $QIR = (E, U, R, RF)$  to a unique, semantically appropriate vertex  $v \in V$  in the Knowledge Graph (KG). Formally, this defines a mapping  $f : E \rightarrow V$ , where each entity is aligned to one KG vertex. To ensure compatibility with arbitrary SPARQL endpoints and avoid KG-specific preprocessing or offline indexing, the Entity Retriever dynamically queries the RDF engine using keyword-based SPARQL. It extracts candidate vertices whose labels contain words from the entity name. The number of candidates per entity is limited by a configurable parameter  $v_{limit}$ . Once the list

<sup>10</sup>Here,  $\text{List}(P)$  denotes a list of predicates, allowing each relation to link to multiple KG predicates rather than a single one.

**Algorithm 2** Vertex and Relation Linking

---

**Input:**  $QIR = (E, U, R, RF)$ : Intermediate representation,  $v_{limit}$ : Maximum number of retrieved vertices,  $endpoint$ : SPARQL endpoint,  $I_{VL}$ : Vertex linking prompt,  $\theta$ : Retry threshold  
**Output:**  $EntToVertex$ : Entity to linked vertices map,  $RelToPred$ : Relation to linked predicates map

```

1:  $EntToVertex, RelToPred \leftarrow \{\}, \{\}$ 
2: for each  $e \in QIR.E$  do
3:    $v_{list} \leftarrow \text{MAgent.entity\_retriever}(e, v_{limit}, endpoint)$ 
4:    $valid \leftarrow \text{False}$ 
5:   while  $valid = \text{False} \wedge \theta > 0$  do
6:      $v_{chosen} \leftarrow \text{MAgent}(e, v_{list}, I_{VL})$ 
7:      $valid \leftarrow \text{MAgent.Validate}(v_{chosen}, v_{list})$ 
8:      $\theta \leftarrow \theta - 1$ 
9:   end while
10:   $EntToVertex[e] \leftarrow v_{chosen}$ 
11: end for
12: for each  $rf \in QIR.RF$  do
13:    $preds \leftarrow \text{MAgent.rel\_Linking}(rf, endpoint, EntToVertex)$ 
14:    $RelToPred[rf.r] \leftarrow preds$ 
15: end for
16: Return  $EntToVertex, RelToPred$ 

```

---

$v_{list} \subseteq V$  is retrieved for an entity  $e$ , the Matching Agent invokes a prompt-guided LLM to select the best candidate:

$$v = f(e, v_{list}, I_{VL}), \quad (5)$$

where  $I_{VL}$  is a handcrafted prompt instructing the LLM to assess label similarity, favoring exact matches when available. To support weaker models, the prompt includes a definition of "exact match." Only vertex labels (not URIs) are provided to improve interpretability, and the output is formatted in JSON for reliable parsing. This zero-shot classification setup allows the LLM to act as a semantic filter under in-context learning, enabling generalization across varied and domain-specific KGs. By decoupling retrieval from selection and avoiding retraining, the Entity Retriever ensures modularity, extensibility, and adaptability without manual schema alignment.

**5.1.2 Vertex Validator.** The Vertex Validator ensures semantic and syntactic correctness of the selected vertex  $v$  before finalizing entity alignment. It enforces two key constraints. First, it verifies that the LLM's output is a syntactically valid and parsable JSON object, suitable for downstream processing. Second, it checks that the selected vertex  $v$  belongs to the retrieved candidate list  $v_{list}$ , guaranteeing that the output corresponds to a real KG entity and not a hallucinated or out-of-scope result. If either validation step fails, the Matching Agent re-invokes the vertex selection tool with a retry budget bounded by the threshold  $\theta$ . This mechanism ensures that only grounded, well-formed vertex selections are propagated to subsequent stages, thereby preserving system robustness and preventing semantic drift in the query generation process.

**5.1.3 Relation Linking.** For relation linking, assigning a single predicate to each semantic relation is often insufficient, as entities in the KG may be connected through multiple semantically related predicates. For example, in DBpedia, the entity *Intel* is linked via

predicates such as *founders*, *founder*, and *foundedBy*. To accurately answer a question like "Who founded Intel?", the system must consider all such alternatives to ensure comprehensive coverage. The Relation Linking tool addresses this by associating each relation  $r \in QIR.R$  with a ranked list of candidate predicates from the KG. For a given relation  $r$ , the tool queries the SPARQL endpoint to retrieve all predicates that connect the previously linked source and target entities. Each predicate URI is converted into a human-readable label by extracting its final segment. Both the relation label and predicate labels are encoded using BERT [9], and their semantic similarity is measured via cosine similarity. The resulting predicates are ranked and stored in the  $RelToPred$  mapping, allowing the downstream query planner to select the most appropriate predicates without relying on KG-specific rules or preprocessing.

**5.1.4 Time and Space Complexity.** The time complexity of the vertex and relation linking phase is  $O(q_{cost} + \theta \cdot C_{LLM})$ , where  $q_{cost}$  is the cost of SPARQL queries used to retrieve candidate vertices and predicates. Since each QIR has only a few entities and relations (typically  $< 10$ ), both SPARQL access and LLM inference operate on bounded input sizes, and validation runs in constant time. The value of  $q_{cost}$  is implementation-dependent and varies with the RDF engine. For space complexity, storing one vertex per entity yields constant overhead for  $EntToVertex$ , while the main cost comes from  $RelToPred$ , which stores ranked predicates. Let  $p_{can}$  be the number of retrieved candidate predicates; total space usage is  $O(p_{can})$ .

## 5.2 Query Planning Agent

Once the Matching Agent produces the mappings of  $EntToVertex$  and  $RelToPred$ , the Query Planning Agent uses them along with the  $QIR$  and the user question  $q_i$ . It manages SPARQL query construction, selection, and execution through three tools: *Query Generator*, *Query Validator*, and *Query Execution*. Query selection is performed directly by the agent. As shown in Lines 1–2 of Algorithm 3, the Query Generator creates a candidate set  $Q = \{\psi_1, \dots, \psi_k\}$  by combining entity mappings from  $EntToVertex$  with predicate options from  $RelToPred$ . To limit latency, the candidate set is truncated to a bounded size  $query\_num$ . The agent then filters  $Q$  based on predicate relevance. Before execution, selected queries are passed to the *Query Validator*, which checks structural correctness and semantic validity. Finally, the validated queries  $Q'$  are executed via the SPARQL endpoint using the Query Execution tool, and the retrieved answers are aggregated into the final result  $A$ , which is returned to the Chat Agent.

**5.2.1 Query Generator.** The Query Generator constructs a candidate set of SPARQL queries  $Q = \{\psi_1, \dots, \psi_k\}$  using the QIR as the structural backbone. It populates each triple  $\langle s, r, o \rangle \in RF$  by replacing entities with their corresponding vertices from  $EntToVertex$ , and substituting relations with predicate options from  $RelToPred(r)$ . Unknowns are included directly as variables. Since each relation may map to multiple predicates, a single semantic triple can yield multiple SPARQL triple patterns. The final queries are formed by joining these patterns based on shared variables or entities. This combinatorial expansion produces multiple candidate queries, each representing a distinct predicate configuration. The main unknown representing the target answer is added to the SELECT clause.

**Algorithm 3** Query Selection and Execution

---

**Input:**  $QIR$ : Intermediate representation,  $EntToVertex$ : Entity-to-vertex map,  $RelToPred$ : Relation-to-predicate map,  $query\_num$ : Query limit,  $q_i$ : User question,  $endpoint$ : SPARQL endpoint,  $I_{QS}$ : Predicate selection prompt,  $\theta$ : Retry threshold  
**Output:**  $A$ : Final answer

```

1:  $Q \leftarrow QPAgent.Gen(QIR, EntToVertex, RelToPred)$ 
2:  $Q \leftarrow Truncate(Q, query\_num)$ 
3:  $\mathcal{P}, PredToQuery \leftarrow [], \{\}$ 
4: for each  $i, \psi \in Q$  do
5:    $P_\psi \leftarrow QPAgent.Pred(\psi)$ 
6:    $\mathcal{P} \leftarrow \mathcal{P} \cup P_\psi$ 
7:   for each  $p \in P_\psi$  do
8:      $PredToQuery[p] \leftarrow PredToQuery[p] \cup \{i\}$ 
9:   end for
10: end for
11:  $valid \leftarrow False, \theta \leftarrow \theta$ 
12: while not valid and  $\theta > 0$  do
13:    $\mathcal{P}' \leftarrow QPAgent(q_i, \mathcal{P}, I_{QS})$ 
14:    $valid \leftarrow QPAgent.Validate(\mathcal{P}', \mathcal{P})$ 
15:    $\theta \leftarrow \theta - 1$ 
16: end while
17:  $Q' \leftarrow \{\psi \in Q \mid QPAgent.Pred(\psi) \cap \mathcal{P}' \neq \emptyset\}$ 
18:  $A \leftarrow QPAgent.Exec(Q', endpoint)$ 
19: return  $A$ 

```

---

**5.2.2 Query Selection.** Lines 3–17 of Algorithm 3 describe the query selection process performed by the Query Planning Agent. The agent iterates over the candidate queries  $\psi \in Q$ , extracts their predicate sets  $Pred(\psi)$ , and aggregates them into a global list  $\mathcal{P}$ . Simultaneously, it builds an inverted index  $PredToQuery$  that maps each predicate to the queries containing it. To reduce execution cost and filter out irrelevant queries, the agent applies an LLM-based predicate filtering strategy. It provides the user question  $q_i$ , predicate list  $\mathcal{P}$ , and a zero-shot chain-of-thought prompt  $I_{QS}$  to the LLM. The model returns a filtered subset  $\mathcal{P}' \subseteq \mathcal{P}$  of predicates considered relevant to the question. This output is validated using the Query Validator to ensure it is well-formed JSON and contains only predicates from the original list. The process is retried up to  $\theta$  times if the output is invalid or malformed. The agent then filters the candidate query set using the validated predicate subset:  $Q' = \{\psi \in Q \mid Pred(\psi) \cap \mathcal{P}' \neq \emptyset\}$ . This step reduces query volume while preserving recall, which help improve efficiency without sacrificing answer accuracy.

**5.2.3 Query Validator.** The Query Validator tool verifies the selected predicates  $\mathcal{P}'$  by ensuring that (i) the LLM output is a well-formed, parsable JSON object, and (ii)  $\mathcal{P}'$  contains only predicates from the candidate set  $\mathcal{P}$ . The validator filters out any predicate in  $\mathcal{P}'$  not found in  $\mathcal{P}$ . If the resulting list is empty, the output is deemed invalid, and the Query Planning Agent retries the selection. Entities are already grounded by the Matching Agent, and SPARQL queries are constructed deterministically from the selected predicates. Therefore, validating  $\mathcal{P}'$  effectively guarantees the correctness of the final queries. These checks mitigate LLM hallucinations with minimal overhead. Since  $\mathcal{P}$  is stored locally as a hash-based structure, validation is efficient.

**5.2.4 Query Execution.** The Query Execution tool dispatches the filtered SPARQL queries  $Q'$  to the knowledge graph and aggregates the results into a unified answer set  $A$ . It interfaces with a standard SPARQL endpoint using API calls, executes each query in  $Q'$ , collects result bindings, removes duplicates, and formats the final output for the Chat Agent. This component is lightweight, stateless, and compatible with standard RDF engines.

**5.2.5 Error Handling and Fallback Strategy.** If no answer is returned, this indicates that the KG may not contain the requested information. The system does not attempt to generate an alternative answer, as this may lead to hallucination. This design ensures the system returns either a factually grounded response based on available data or explicitly indicates that the requested information is not present in the KG. Inconsistent results are avoided through assertion-based validations applied throughout the pipeline, especially during query selection. These validations ensure that intermediate outputs remain relevant to the user's question.

**Time and Space Complexity** Algorithm 3 has a time complexity of  $O(q_{cost} + \theta \cdot C_{LLM})$ , where  $q_{cost}$  denotes the cost of SPARQL query execution and  $C_{LLM}$  is the cost of a single LLM call. The loop over candidate queries (Lines 3–10) is bounded by the truncation parameter  $query\_num$ , typically less than 100, and does not affect asymptotic complexity. LLM output validation (Lines 11–16) incurs constant time per attempt, with up to  $\theta$  retries. Space complexity is dominated by the storage of candidate queries and the predicate-to-query index. Both scale linearly with  $query\_num$ , resulting in overall space complexity of  $O(query\_num)$ . Since fewer queries are usually executed than  $query\_num$ , the algorithm keeps overhead low while preserving accuracy and robustness.

## 6 Implementation

We implement CHATTY-KG<sup>11</sup> using the LangGraph framework<sup>12</sup>, an extension of LangChain that introduces graph-based abstractions for multi-agent workflows. LangGraph enables structured control over agent composition and inter-agent communication via a shared state and programmable routing logic. Each agent in CHATTY-KG is implemented as an independent function that reads and updates a shared AgentState object. This object acts as centralized memory, storing the user question, intermediate outputs, and final results. These agent functions are composed into a StateGraph, where conditional transitions depend on the evolving state. As shown in Listing 1, execution begins at `chat_agent` (Line 19) and proceeds through agent nodes (Lines 12–17), including `classifier_agent`, `rephraser_agent`, and `qir_agent`. While Figure 2 depicts a fixed logical view, the implementation is dynamic: each agent sets the route field to determine the next node. This decouples routing from static topology, enabling context-sensitive transitions and modular reconfiguration without altering control flow.

The modular design of CHATTY-KG enables adaptability at multiple levels. **LLM substitution** is straightforward: agents can use different underlying models (e.g., GPT-4, Phi, Mistral) by updating internal configurations. **Agent extensibility** is also supported, allowing components to be swapped or extended, e.g., replacing a

<sup>11</sup><https://github.com/CoDS-GCS/Chatty-KG>

<sup>12</sup><https://python.langchain.com/docs/langgraph/>



**Listing 1: LangGraph-based implementation of CHATTY-KG. The shared agent state tracks intermediate outputs, such as rephrased questions, SPARQL queries, and routing decisions, exchanged among agents during dialogue processing.**

```

1 # Agent function signatures (each takes and returns AgentState)
2 def chat_agent(state: AgentState) -> AgentState: ...
3 def classifier_agent(state: AgentState) -> AgentState: ...
4 def rephraser_agent(state: AgentState) -> AgentState: ...
5 def qir_agent(state: AgentState) -> AgentState: ...
6 def query_agent(state: AgentState) -> AgentState: ...
7 def matching_agent(state: AgentState) -> AgentState: ...
8
9 # Graph construction
10 builder = StateGraph(state_schema=AgentState)
11
12 builder.add_node("chat_agent", chat_agent)
13 builder.add_node("classifier_agent", classifier_agent)
14 builder.add_node("rephraser_agent", rephraser_agent)
15 builder.add_node("qir_agent", qir_agent)
16 builder.add_node("query_agent", query_agent)
17 builder.add_node("matching_agent", matching_agent)
18
19 builder.set_entry_point("chat_agent")
20
21 # Define transitions between agent nodes
22 builder.add_edge("chat_agent",
23     lambda state: END if state.query_done else ("query_agent" if
24         state.qir_done else "qir_agent"))
25 builder.add_conditional_edges("classifier_agent",
26     lambda state: state.route)
27 builder.add_conditional_edges("rephraser_agent",
28     lambda state: state.route)
29 builder.add_conditional_edges("qir_agent",
30     lambda state: state.route)
31 builder.add_conditional_edges("query_agent",
32     lambda state: state.route)
33 builder.add_conditional_edges("matching_agent",
34     lambda state: state.route)
35
36 builder.set_finish_point("chat_agent")
37 graph = builder.compile()

```

rule-based tool with a neural module, without retraining or reengineering the system. Finally, **dialogue mode control** is governed by a `system_mode` flag within the shared state, allowing the system to toggle between multi-turn and single-turn execution. These features make CHATTY-KG a flexible and extensible platform for research and deployment in diverse KGQA settings.

CHATTY-KG uses four main parameters: (1)  $\theta$ , the maximum number of retries for LLM calls when validation fails; (2)  $L$ , the number of past answers included in the context  $C_q$  to fit within the LLM’s input length; (3)  $v_{limit}$ , the number of candidate vertices retrieved by the entity retriever; and (4)  $query\_num$ , the maximum number of candidate queries considered during query selection. In our experiments, we set these to 3, 100, 600, and 40, respectively. The  $\theta$  parameter helps reduce hallucinations by allowing the LLM multiple attempts to produce valid outputs. While higher values improve robustness, they also increase cost and latency. Models like GPT-4o and Gemini benefit from retries by producing varied and often valid responses, whereas weaker models like Llama show limited gains. We set  $L=100$  based on empirical testing. Larger values may exceed the LLM context window for long answers (e.g., lists of papers), while smaller values risk omitting context needed by the rephraser. The  $v_{limit}$  parameter balances recall and prompt complexity: higher values improve recall but add noise, while lower values reduce noise but risk missing the correct vertex. The  $query\_num$  value must retain valid queries without increasing runtime. For fairness in evaluation, we disable natural-language reformulation and compare against structured gold-standard answers.

**Table 2: KG Statistics: The number of triples and entities in Millions, and the number of predicates of each KG.**

KG	# Entities(M)	# Triples(M)	# Predicates
DBLP	18	263	167
YAGO	12	207	259
DBpedia	14	350	60,736
MAG	586	13,705	178
Wikidata <sup>13</sup>	119	8,541	12,943

## 7 Evaluation

We evaluate CHATTY-KG along five key dimensions: (1) performance on single-turn QA compared to state-of-the-art KGQA systems, (2) accuracy on multi-turn questions over KGs relative to top-performing chatbots, LLMs and RAG-based systems, (3) the contribution of our contextual understanding module, (4) the effect of our LLM-powered question interpretation, and (5) the impact of our query planning and selection on system efficiency.

### 7.1 Evaluation Setup

**Baselines:** We evaluate CHATTY-KG against state-of-the-art systems in two settings: single-turn and multi-turn question answering. For single-turn, we compare against KGQA [30], a leading KGQA system, and EDGQA [15]. For multi-turn (conversational) QA, we compare against CONVINSE [5] and EXPLAIGNN [6]. Both are designed for KG-based dialogue and trained on the ConvMix dataset [5], which augments multi-turn QA over Wikidata. Unlike CHATTY-KG, these systems require task-specific training and are tightly coupled to the KG they were trained on. We evaluated two widely used **RAG systems**, ColBERT [21, 37] and GraphRAG [13], on DBLP (our smallest KG). Both require substantial compute and memory. ColBERT performs late-interaction retrieval without building a graph, while GraphRAG constructs and indexes one for downstream QA. Indexing DBLP with GraphRAG required over 80 hours using a 512 GB, 64-core VM, used solely for the RAG experiment. Extrapolating from this, larger KGs would require terabytes of memory, making RAG indexing impractical in our setting.

**Underlying LLMs:** We evaluate CHATTY-KG using: (1) Commercial LLMs, GPT-4o [32], Gemini-2.0-Flash [42], and DeepSeek-V3 [8], accessed via API. (2) Open-weight LLMs, such as Qwen-2.5 [51], Phi-4 [1], Gemma-3 [19], CodeLlama [35], Gemma-2 [34], Granite [16], Vicuna [27], Llama-3 [2], and Mistral [18], all hosted locally.

**Five Real-World KGs:** To assess CHATTY-KG’s performance across domains, we evaluate it on five real-world KGs: DBpedia [24], YAGO [11, 41], DBLP [7], Microsoft Academic Graph (MAG) [28], and Wikidata [46]. DBpedia, Wikidata and YAGO cover general knowledge (e.g., people, places), while DBLP and MAG focus on academic content, including long entity names like paper titles. These KGs vary in size, enabling evaluation of CHATTY-KG’s scalability and robustness. Table 2 summarizes the number of entities, predicates, and triples. For wikidata SPARQL execution, we use the public available endpoint. For the other graph’s SPARQL execution, we use Virtuoso v7.2.5.2, a standard backend for large KGs, with a separate endpoint per KG. All experiments use the default Virtuoso setup.

**Compute Infrastructure:** We use two different setups for our experiments: (i) Linux machine with 16 cores and 180GB RAM for

<sup>13</sup><https://www.wikidata.org/wiki/Wikidata:Statistics>

**Table 3: Comparison between state-of-the-art KGQA systems and CHATTY-KG on single-turn question answering across five benchmarks covering four knowledge graphs. CHATTY-KG is evaluated with various LLMs. Open-weight LLMs are ordered top-down by parameter count. Bold indicates best performance, underlined indicates second-best. Green highlights performance better than KGQA systems; yellow highlights results within 1 point. RAG systems require huge memory to run.**

System(Model)	QALD-9			LC-QuAD 1.0			YAGO			DBLP			MAG		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
KGQAN	51.13	38.72	44.07	58.71	46.11	51.65	48.48	65.22	55.62	57.87	52.02	54.79	55.43	45.61	50.05
EDGQA	31.30	40.30	32.00	50.50	<b>56.00</b>	53.10	41.90	40.80	41.40	8.00	8.00	8.00	4.00	4.00	4.00
GraphRAG (GPT-4o)	-	-	-	-	-	-	-	-	-	46.00	37.01	38.99	-	-	-
COLBERT (GPT-4o)	-	-	-	-	-	-	-	-	-	65.38	55.33	55.29	-	-	-
CHATTY-KG(GPT-4o)	<b>61.91</b>	41.50	<b>49.69</b>	<b>72.79</b>	43.60	<b>54.54</b>	86.52	72.27	<b>78.75</b>	85.96	75.50	<b>80.39</b>	<b>75.05</b>	69.05	<b>71.92</b>
CHATTY-KG(Gemini-2.0-Flash)	<u>57.23</u>	<b>42.14</b>	<b>48.54</b>	68.27	43.35	<b>53.03</b>	<b>90.04</b>	<b>75.47</b>	<b>82.11</b>	74.80	67.50	70.96	59.55	57.56	58.54
CHATTY-KG(DeepSeek-V3)	50.76	40.83	<b>45.26</b>	<u>69.66</u>	45.42	<b>54.99</b>	78.77	<u>74.07</u>	<b>76.34</b>	71.88	<b>86.50</b>	<b>78.52</b>	64.43	<b>72.36</b>	<b>68.16</b>
CHATTY-KG(Phi-4)	55.95	37.75	<b>45.08</b>	69.09	42.43	<b>52.57</b>	<b>86.89</b>	<b>74.07</b>	<b>79.97</b>	81.04	75.50	<b>78.17</b>	63.23	62.70	<b>62.96</b>
CHATTY-KG(Qwen2.5-14B-Inst)	50.77	42.17	<b>46.07</b>	68.84	42.69	<b>52.70</b>	<b>83.73</b>	<b>67.07</b>	<b>74.48</b>	65.96	73.50	<b>69.53</b>	57.58	61.93	59.68
CHATTY-KG(Qwen2.5-14B)	48.99	39.84	<b>43.94</b>	62.49	41.29	49.72	74.77	71.76	<b>73.23</b>	58.61	68.50	<b>63.17</b>	61.73	62.56	<b>62.14</b>
CHATTY-KG(Vicuna-13B-v1.5)	62.40	26.17	36.87	73.78	27.92	40.51	58.53	36.00	44.58	64.90	30.50	41.50	75.93	34.51	47.45
CHATTY-KG(CodeLlama-13B)	43.99	32.32	37.26	56.31	38.41	45.67	44.48	53.24	48.47	39.08	68.00	49.64	34.86	61.82	44.58
CHATTY-KG(Gemma3-Inst)	53.64	36.93	<b>43.75</b>	66.13	39.09	49.13	72.67	65.87	<b>69.10</b>	71.44	76.50	<b>73.88</b>	49.28	43.52	46.22
CHATTY-KG(Gemma3)	57.11	34.76	<b>43.22</b>	63.65	39.33	48.62	75.33	72.79	<b>74.04</b>	57.06	<u>63.50</u>	<b>60.11</b>	54.64	51.30	<b>52.92</b>
CHATTY-KG(Gemma2-Inst)	51.16	35.47	41.90	65.88	39.85	49.66	81.57	60.47	<b>69.45</b>	79.81	68.00	<b>73.43</b>	<b>65.95</b>	49.22	<b>56.37</b>
CHATTY-KG(Gemma2)	43.19	37.59	40.20	53.95	41.56	46.95	62.93	66.47	<b>64.65</b>	48.46	58.00	52.80	48.19	44.23	46.12
CHATTY-KG(Ministral-8B-Inst)	56.72	37.47	<b>45.13</b>	65.31	39.98	49.60	78.92	72.19	<b>75.41</b>	<b>87.94</b>	69.50	<b>77.64</b>	58.17	58.54	<b>58.35</b>
CHATTY-KG(Granite-3.2-8B-Inst)	47.23	36.68	41.29	66.59	34.27	45.25	69.68	72.46	71.04	74.09	63.00	<b>68.10</b>	51.00	56.71	<b>53.70</b>
CHATTY-KG(Llama-3.1-8B-Inst)	56.31	32.54	41.25	65.66	37.63	47.84	67.61	70.66	<b>69.10</b>	66.24	60.01	<b>62.98</b>	63.87	64.06	<b>63.97</b>
CHATTY-KG(Mistral-7B-v0.3)	48.75	36.02	41.43	61.39	33.41	43.27	50.02	65.99	<b>56.91</b>	47.36	63.00	<b>54.07</b>	35.09	59.28	44.09

running all experiments on CHATTY-KG and KGQAN. It is also used to host the four Virtuoso SPARQL endpoints for the KGs. (ii) Linux machine with Nvidia A100 40GB GPU, 32 cores, and 64GB RAM to host the open-weight LLMs, creating a server accessible using HTTP requests via the VLLM library [23].

**Benchmarks:** We evaluate CHATTY-KG on a comprehensive set of KGQA benchmarks spanning both single-turn and multi-turn QA over multiple KGs. For **single-turn QA**, we use standard datasets such as QALD-9 [44] and LC-QuAD 1.0 [43], which provide SPARQL-annotated questions and gold answers over DBpedia. QALD-9 includes 408 training and 150 test questions, while LC-QuAD 1.0 provides 4,000 training and 1,000 test questions generated from diverse templates. Since CHATTY-KG requires no training, we only use the test sets. We also include datasets introduced by [30], with 100 questions each for YAGO, DBLP, and MAG, covering varied domains and KG structures. For **multi-turn QA**, we generate 20 dialogues per KG (approximately 100 questions) using a state-of-the-art dialogue benchmark generator [31]. Each dialogue consists of 5 turns, with every question  $q$  paired with its standalone version, a corresponding SPARQL query, and the correct answer. These generated benchmarks allow us to evaluate CHATTY-KG’s contextual understanding and reasoning capabilities in a reproducible setting.

**Metrics:** For single-turn QA, we use standard KGQA metrics: Precision (P), Recall (R), and F1 score. Recall measures the proportion of correct answers returned by the system, while Precision measures the proportion of returned answers that are correct. The F1 score is the harmonic mean of Precision and Recall. For multi-turn QA, we adopt the metrics used by CONVINSE and EXPLAINNN: Precision at 1 (P@1), Mean Reciprocal Rank (MRR), and Hit at 5 (Hit@5). P@1 checks whether the top-ranked answer is correct and is common when systems return a single answer [5]. MRR evaluates the position of the correct answer in a ranked list. Hit@5 measures whether the correct answer appears among the top five results.

## 7.2 Single-Turn KGQA

We evaluate CHATTY-KG on single-turn QA using five benchmarks and compare it with two state-of-the-art systems: KGQAN and EDGQA. Results are reported in Table 3. We also assess how CHATTY-KG’s performance varies when paired with different commercial and open-weight LLMs. CHATTY-KG outperforms both baselines when using high-performing LLMs like GPT-4o, Gemini-2.0, and DeepSeek. For example, GPT-4o achieves F1 scores of 80.39 on DBLP and 71.92 on MAG, compared to KGQAN’s 54.79 and 50.05. On QALD and LC-QuAD, it leads with F1 scores of 49.69 and 54.54, improving over KGQAN by +5.62 and +2.89, respectively.

GPT-4o also yields strong precision, such as 86.52 on YAGO, showing that CHATTY-KG’s query planning effectively filters incorrect queries. DeepSeek-V3 achieves high recall, e.g., 86.50 on DBLP. Models like Phi-4 and Qwen2.5-14B-Instruct provide a good trade-off between precision and recall, outperforming KGQAN in the five benchmarks. Notably, even smaller open-weight models like Mistral-7B and Llama-3.1-8B match or exceed KGQAN on some datasets. This shows that CHATTY-KG achieves strong performance without task-specific fine-tuning. Our modular prompting and accurate query selection enable this capability, even with smaller models.

EDGQA performs competitively on LC-QuAD (F1 = 53.10) but fails on domain-specific KGs, such as DBLP (F1 = 8.0) and MAG (F1 = 4.0), likely due to its rule-based architecture. KGQAN performs better overall but struggles with precision on general datasets like QALD-9 (P = 51.13) and YAGO (P = 48.48), often retrieving irrelevant answers. In contrast, CHATTY-KG maintains high precision across both general and special domains. It delivers strong F1 and precision scores on datasets like DBLP and YAGO. These results demonstrate the effectiveness of our query planning agent. Given GPT-4o’s consistent performance across all datasets and metrics, we use GPT-4o as the default model for the remainder of our evaluation, unless otherwise specified.

**Table 4: Performance of CHATTY-KG on multi-turn questions, compared to conversational baselines (CONVINSE, EXPLAIGNN) and general LLMs (GPT-4o, Gemini, DeepSeek, Phi-4, Qwen) across three KGs. Metrics: P@1, MRR, and Hit@5. CHATTY-KG outperforms all baselines, achieving up to +87% higher P@1 on YAGO and nearly 3× higher accuracy on DBLP.**

System	Wikidata			DBpedia			YAGO			DBLP		
	P@1	MRR	Hit@5	P@1	MRR	Hit@5	P@1	MRR	Hit@5	P@1	MRR	Hit@5
CONVINSE	27.47	27.47	27.47	28.72	28.72	28.72	34.88	34.88	34.88	28.89	28.89	28.89
EXPLAIGNN	26.37	26.37	26.37	29.79	29.79	29.79	31.40	31.40	31.40	28.89	28.89	28.89
GraphRAG (GPT-4o)	-	-	-	-	-	-	-	-	-	11.10	11.10	11.10
ColBERT (GPT-4o)	-	-	-	-	-	-	-	-	-	17.78	17.78	17.78
GPT-4o	49.45	49.45	49.45	30.85	30.85	30.85	27.91	27.91	30.23	24.44	24.44	24.44
Gemini-2.0-Flash	41.76	41.76	41.76	25.53	24.47	24.47	30.23	31.10	32.56	13.33	13.33	13.33
DeepSeek-V3	40.65	40.65	40.65	20.21	20.21	20.21	26.74	26.74	26.74	16.67	16.67	16.67
Phi-4	25.27	25.27	25.27	26.60	25.53	25.53	15.12	15.31	15.12	17.78	17.78	17.78
Qwen2.5-14B-Instruct	24.18	24.18	24.18	18.09	18.09	18.09	16.28	16.86	17.44	6.67	5.56	5.56
CHATTY-KG(GPT-4o)	54.95	54.95	54.95	34.04	35.11	36.17	65.12	65.70	66.28	83.33	83.33	83.33

We also evaluated two RAG systems, GraphRAG and ColBERT, to contextualize our results. On DBLP, GraphRAG achieved moderate precision (46.00) but low recall (37.01), while ColBERT performed better ( $P = 65.38$ ,  $R = 55.33$ ,  $F1 = 55.29$ ). Both struggled with list-style questions, often returning only partial answers, and failed on multi-hop queries because text chunking breaks graph structure. GraphRAG also required heavy indexing and still missed relations, confirming the scalability challenge. These results support our observation that current RAG implementations cannot reliably preserve structure or guarantee complete answers. In contrast, CHATTY-KG operates directly over the KG and executes SPARQL, enabling accurate, complete retrieval and consistent performance.

### 7.3 Multi-turn Dialogues

We evaluate CHATTY-KG on multi-turn question answering using 20 dialogues (up to 100 questions) generated by a state-of-the-art benchmark generator [31]. We compare against two conversational KGQA systems, CONVINSE and EXPLAIGNN, and several general-purpose LLMs that do not have access to KGs (GPT-4o, Gemini-2.0-Flash, DeepSeek-V3, Phi-4, and Qwen2.5-14B-Instruct).

**Dialogue Datasets.** CONVINSE and EXPLAIGNN were trained on Wikidata, while we evaluate on DBpedia, YAGO, and DBLP. To ensure fairness, we manually selected 20 entities that exist in both Wikidata and our target KGs. For DBpedia and YAGO, we verified alignment using Wikidata links. For DBLP, which lacks clear mapping, we selected entities relevant to research discussions (e.g., “Gerhard Kramer”). CONVINSE and EXPLAIGNN were accessed through their official demos<sup>14,15</sup>, and their outputs were converted to match our evaluation format. The performance of CONVINSE and EXPLAIGNN in our setup matches the results reported in their original papers, confirming the fairness of our experimental design.

**Performance Results.** Table 4 presents results using P@1, MRR, and Hit@5. CHATTY-KG outperforms all baselines and LLMs across every KG and metric. On Wikidata, it achieves 54.95 P@1, significantly outperforming CONVINSE (27.47) and EXPLAIGNN (26.37) which were trained on Wikidata. On YAGO, it achieves an 87%

**Table 5: Performance comparison of CHATTY-KG with and without the conversational module using GPT-4o. Dialogue results reflect context-dependent questions handled with the module, while Standalone results use context-free questions. Retention (%) =  $F1_{\text{Dialogue}} / F1_{\text{Standalone}} \times 100$  indicates the proportion of accuracy retained in dialogue settings.**

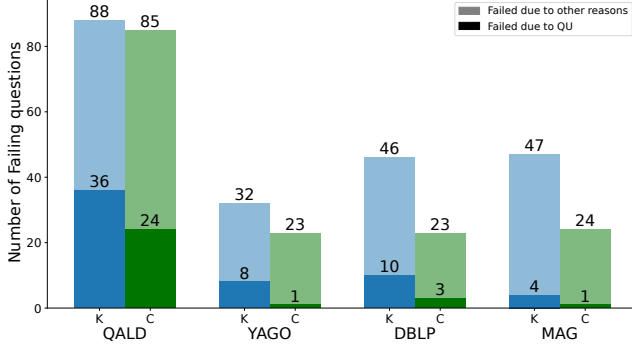
Dataset	Method	P	R	F1	Retention (%)
Wikidata	Dialogue	78.65	57.14	66.19	95.28
	Standalone	79.75	61.54	69.47	
DBpedia	Dialogue	68.19	42.55	52.40	92.27
	Standalone	69.79	47.87	56.79	
YAGO	Dialogue	74.86	67.44	70.96	85.48
	Standalone	84.69	81.40	83.01	
DBLP	Dialogue	81.85	74.44	77.97	87.09
	Standalone	89.07	90.00	89.53	

performance gain in P@1 compared to CONVINSE. On DBLP, a domain-specific KG, it reaches 83.33 P@1, nearly 3X the score of the best baseline. On DBpedia, CHATTY-KG still leads with 34.04 P@1, ahead of EXPLAIGNN’s 29.79. The improvement on DBpedia is smaller due to its scale and structure. As shown in Table 2, DBpedia has 60,736 predicates, about 234× more than YAGO and 4× more than Wikidata, which introduces high ambiguity in predicate selection. Many predicates are also redundant or synonymous (e.g., *founders*, *founder*, *foundedBy* for *Intel*), making query planning more difficult. Despite this, CHATTY-KG remains competitive on DBpedia and does so without KG-specific training or preprocessing. Our modular agents ground responses in the KG through accurate linking, context reasoning, and minimal query execution. This design adapts to different graph structures and remains robust across domains without retraining or pipeline changes.

We also evaluated two RAG systems, GraphRAG and ColBERT, in the same multi-turn setting. GraphRAG performed poorly (e.g.,  $P@1 = 11.10$  on DBLP), and ColBERT showed similarly low results ( $P@1 = 17.78$ ). Both systems handle each query independently and lack dialogue-state tracking, so they cannot resolve references or

<sup>14</sup><https://convinse.mpi-inf.mpg.de/>

<sup>15</sup><https://explainnn.mpi-inf.mpg.de/>



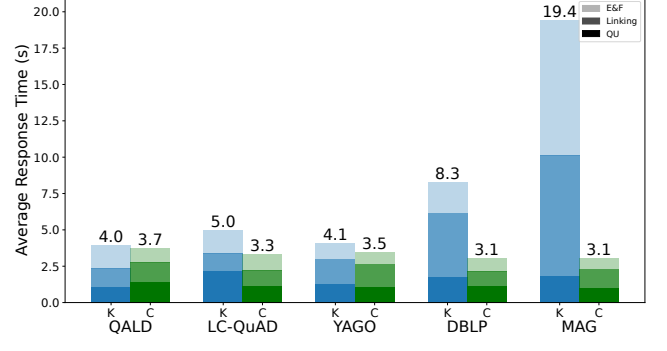
**Figure 3: Number of failed questions (i.e., Recall = 0) in each benchmark. Each bar is divided to show failures caused by Question Understanding (QU) and other factors. Shading variations indicate the source of failure. CHATTY-KG consistently has the fewest failures across all benchmarks.**

preserve context across turns. Their retrieval resets at every step, breaking conversational coherence and leading to low accuracy. In contrast, CHATTY-KG maintains dialogue context and delivers consistent multi-turn performance.

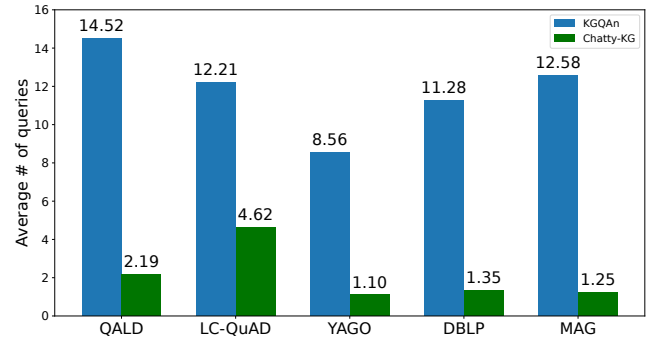
**General LLM Limitations.** While LLMs used without KG access like GPT-4o and Gemini show moderate performance on DBpedia and YAGO, they struggle on DBLP. For instance, GPT-4o drops to 24.44 P@1 on DBLP, far behind CHATTY-KG. This is expected, as domain-specific KGs like DBLP may be underrepresented in LLM training data. In contrast, CHATTY-KG grounds its answers directly in the KG and executes structured queries. This design enables higher accuracy, better verification, and lower latency. It handles evolving domains without retraining or KG-specific pipelines. As LLMs are limited by static training data, CHATTY-KG provides a more reliable and up-to-date solution for multi-turn KGQA. Its combination of contextual understanding and precise query planning ensures high accuracy and low latency without requiring retraining.

## 7.4 Analyzing CHATTY-KG

**Contextual Understanding:** We evaluate CHATTY-KG’s ability to handle multi-turn questions. For this, we reuse the same benchmarks from the multi-turn evaluation with CONVINSE and EXPLAIGNN. Each dialogue in these benchmarks includes aligned question pairs: a context-dependent version and its corresponding standalone form. We run CHATTY-KG in two modes: (1) with the Classifier and Rephraser Agents enabled, using the context-dependent questions; and (2) with these agents disabled, using the standalone versions. This setup tests the system’s ability to reconstruct missing context from dialogue history. As shown in Table 5, CHATTY-KG retains over 85% of its standalone F1 score across all datasets. This demonstrates the effectiveness of the Classifier and Rephraser Agents in resolving contextual dependencies with minimal performance loss. The slight drop reflects the inherent ambiguity in multi-turn inputs, which CHATTY-KG mitigates effectively. **Question Understanding:** We evaluate CHATTY-KG’s ability to understand questions by analyzing failure cases, as shown in Figure 3. The figure compares the number of failing questions (Recall=0) for CHATTY-KG (denoted as C) and KQAN (denoted as K) across four benchmarks: QALD, YAGO, DBLP, and MAG. Each bar is split into



**Figure 4: Average response time per question (in seconds) for KQAN (K) and CHATTY-KG (C). Each bar is segmented bottom-up into three stages: Question Understanding (QU), Linking, and Execution & Filtration (E&F). Shading variations within each bar indicate the contribution of each stage.**



**Figure 5: Average number of queries per question for KQAN and CHATTY-KG. Lower values indicate better query planning, while also CHATTY-KG achieves higher F1, see Table 3.**

two segments: failures due to **question understanding (QU)** and failures due to **other reasons**. Darker shades indicate QU-related failures, while lighter shades capture all other causes. Across all benchmarks, CHATTY-KG (C) shows significantly fewer failures than KQAN (K), especially in question understanding. For example, in the YAGO and MAG datasets, CHATTY-KG fails due to QU in only one case, compared to eight and four cases, respectively, for KQAN. This demonstrates that CHATTY-KG’s LLM-based modular agents are more robust in interpreting diverse and complex linguistic questions, outperforming KQAN’s fine-tuned model.

Improved question understanding helps downstream accuracy, but gains are not always linear because errors can still happen in later stages. The remaining errors (lighter shades in Figure 3) come from structural mismatches, entity linking, and query selection. In QALD-9, many failures stem from mismatches between natural language and KG structure. For example, “How many emperors did China have?” should link to *Emperor of China*, not *China*. Likewise, “Who killed Caesar?” should retrieve *Assassins of Julius Caesar*, not just *Caesar*. Some questions also rely on vague KG predicates (e.g., *subject*), which do not clearly capture the intended semantics. In these cases, question parsing is correct, but the KG organizes knowledge differently than natural language. Entity linking errors occur when the model chooses the wrong candidate due to ambiguity or similar entity names. For example, in “Are Taiko some kind of Japanese musical instrument?”, the system links to <http://>



**Table 6: Results of running CHATTY-KG with the translation extension module on the 7 languages of QALD-9.**

Lang	en	de	pt	hi	fr	nl	it
P	62.29	66.31	63.62	65.20	64.21	65.80	62.97
R	40.83	42.16	42.14	39.36	37.72	36.40	37.14
F1	49.33	50.79	50.70	49.08	47.52	46.87	46.72

**Table 7: Human and model evaluation across quality dimensions, including inter-annotator agreement.**

Metric	ChatGPT	Gemini	H1	H2	H3	Avg	Std
Response Quality	4.48	4.64	4.50	4.43	4.22	4.45	0.15
Fluency	4.92	4.86	4.88	4.76	4.86	4.86	0.06
Dialogue Coherence	4.52	4.52	4.50	4.45	4.57	4.51	0.04
Weighted Cohen's $\kappa$ (Human-Human):			0.53 (Moderate Agreement)				
Weighted Cohen's $\kappa$ (Human-LLM):			0.56 (Moderate Agreement)				

dbpedia.org/resource/Japanese\_musical\_instrument instead of <http://dbpedia.org/class/yago/WikicatJapaneseMusicalInstruments>.

Query selection errors arise when the agent either rejects all queries or picks one with the wrong predicate. For instance, in “*How many grand-children did Jacques Cousteau have?*”, it selects *relative* instead of *child*. These errors can cascade: a wrong entity link weakens query candidates, causing incorrect rejection or ranking. Chatty-KG retries up to three times to reduce such failures while controlling latency and cost. Despite these challenges, it still cuts both overall errors and QU-specific errors by a large margin.

**System Efficiency:** We evaluate CHATTY-KG’s runtime efficiency against KGQAN based on average response time and average number of executed queries. Both are critical for interactive systems where fast and precise answers are expected. Results are shown in Figures 4 and 5. Figure 4 breaks down response time into three stages: Question Understanding (QU), Linking, and Execution & Filtration (E&F). CHATTY-KG (C) consistently outperforms KGQAN (K) in all stages, with the largest gains in the E&F phase. For example, on MAG, KGQAN takes 19.4 seconds per question, while CHATTY-KG completes in just 3.1 seconds. This demonstrates CHATTY-KG’s scalability to large graphs. Figure 5 shows that CHATTY-KG executes up to 10× fewer queries than KGQAN. This reduction lowers response time and improves precision by avoiding irrelevant queries, as shown in Figures 4 and 5. Beyond response time, we also measured cost on QALD-9 (150 questions). Running CHATTY-KG with GPT-4o used 507 requests, 326K input tokens, and 7.73K output tokens, for a total cost of \$0.88. This shows that CHATTY-KG remains cost-effective even with multiple LLM calls.

## 7.5 Analyzing CHATTY-KG and Discussion

**Multilingual Support** is enabled by adding an LLM translation module to the Chat Agent, which converts non-English questions to English before processing. This allows a single pipeline, since the KG and SPARQL entities are in English. On QALD-9, CHATTY-KG yields consistent scores across 7 languages (Table 6), with F1 ranging from 46.72 (it) to 50.79 (de) vs. 49.33 (en). Some languages outperform English as translation often clarifies entity mentions, improving linking, while lower scores reflect translation quality variance. We analyzed 11 languages in total<sup>16</sup>.

<sup>16</sup>See [supplementary materials](#) for more details and examples.

**Table 8: Number of questions solved by CHATTY-KG across different question types in KGQA benchmarks. Results are shown as system solved/total questions format.**

Benchmark	Factoid	Count	Boolean	Q Len	Preds
QALD-9	63/133	2/13	0/4	7.52	1.77
YAGO-B	77/99	-	0/1	6.97	2.00
DBLP-B	70/92	7/8	-	9.73	1.72
MAG-B	58/81	16/17	2/2	8.63	1.87
DBpedia-Dia	34/64	0/11	6/19	7.10	1.00
YAGO-Dia	48/65	0/3	10/18	8.21	1.02
DBLP-Dia	48/59	2/4	17/27	11.62	1.01
Wikidata-Dia	36/61	1/5	15/25	9.05	1.03

**Human and LLM Evaluation:** To assess conversational quality, we used **five evaluators**: three human annotators and two LLMs (CHATGPT, GEMINI).<sup>17</sup> Evaluators rated each response on *Response Quality* and *Fluency*, and scored the full dialogue for *Dialogue Coherence*. All scores used a 1–5 scale. As shown in Table 7, CHATTY-KG achieved strong results: **4.45** in Response Quality, **4.86** in Fluency, and **4.51** in Dialogue Coherence, with low variance. Human-human agreement reached  $\kappa = 0.53$ , and human-LLM agreement  $\kappa = 0.56$ , indicating *moderate agreement*. This reflects consistent judgments across annotators and reliable quality ratings.

**Question Type Analysis:** KGQA systems are best for fact-based queries (who, what, where, when, count) and are less suitable for open-ended ones that need external reasoning (why, how). We group questions into three types: **Factoid** (entity or literal answers), **Count** (requires counting), and **Boolean** (yes/no verification).<sup>18</sup> Table 8 reports CHATTY-KG’s performance across these types. Factoid questions dominate benchmarks and CHATTY-KG solves most of them across datasets. Count and Boolean accuracy varies by dataset, as these tasks are more sensitive to query structure and KG coverage. We report average question length (no. of words, **Q Len**) and average predicates per query (**Preds**). Dialogue datasets use fewer predicates since context carries across turns, reducing the need to explicitly specify all relations in each query.

**Training-Free Prompting:** Chatty-KG uses task-aware prompting rather than training. Simple agents use zero-shot prompts (e.g., Rephraser), while more complex agents use few-shot prompts (e.g., QIR) and chain-of-thought when needed. This avoids dataset creation and fine-tuning, enabling fast deployment. Although light fine-tuning could help, it requires task- and KG-specific data for each module, reducing flexibility on new or evolving graphs. Empirically, prompting alone outperforms KGQAN (which fine-tunes for question understanding) as shown in Figure 3. Prompting offers scalability, cross-KG generalization, and practical deployment without specialized training data.

## 8 Conclusion

Conversational QA over KGs enables accurate and timely access to enterprise and domain knowledge. Existing systems remain limited, mostly single-turn and dependent on heavy preprocessing or fine-tuning. General LLMs also lack grounding in structured data,

<sup>17</sup>Full protocol and examples are in the [supplementary materials](#).

<sup>18</sup>More details are in the [supplementary materials](#).



reducing reliability on private or evolving graphs. CHATTY-KG combines structured reasoning with LLM dialogue through a modular multi-agent design for multi-turn KGQA. It uses LLM-powered agents for context tracking, linking, and query planning without KG-specific training, enabling domain transfer, coherent dialogue, and efficient SPARQL generation across diverse graphs. **Key learnings:** (1) Modular agents allow incremental improvements without system-wide changes. (2) A shared state enables coordinated execution. (3) Agent-level validation reduces hallucination by catching errors early. (4) Bounded retries prevent infinite loops when tasks exceed LLM capability. (5) The multi-agent design adds little latency and debugging becomes more complex. These choices deliver accuracy, adaptability, and practical performance for real-world KGQA.

## References

- [1] Marah I Abdin, Jyoti Aneja, Harkirat S. Behl, Sébastien Bubeck, Ronen Eldan, and et al. 2024. Phi-4 Technical Report. *CoRR* abs/2412.08905 (2024). arXiv:2412.08905 <https://doi.org/10.48550/arXiv.2412.08905>
- [2] AI@Meta. 2024. Llama 3.1 Model Card. [https://github.com/meta-llama/llama-models/blob/main/models/llama3\\_1/MODEL\\_CARD.md](https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md)
- [3] Diego Ceccarelli, Claudio Lucchese, Salvatore Orlando, Raffaele Perego, and Salvatore Trani. 2014. Dexter 2.0 - an Open Source Tool for Semantically Enriching Data. In *Proceedings of the International Semantic Web Conference, Posters & Demonstrations Track (ISWC)*, Vol. 1272. 417–420. [http://ceur-ws.org/Vol-1272/paper\\_127.pdf](http://ceur-ws.org/Vol-1272/paper_127.pdf)
- [4] Philipp Christmann, Rishiraj Saha Roy, and Gerhard Weikum. 2022. Beyond NED: Fast and Effective Search Space Reduction for Complex Question Answering over Knowledge Bases. In *WSDM: The ACM International Conference on Web Search and Data Mining*. 172–180. <https://doi.org/10.1145/3488560.3498488>
- [5] Philipp Christmann, Rishiraj Saha Roy, and Gerhard Weikum. 2022. Conversational Question Answering on Heterogeneous Sources. In *SIGIR: The International ACM SIGIR Conference on Research and Development in Information Retrieval*. 144–154. <https://doi.org/10.1145/3477495.3531815>
- [6] Philipp Christmann, Rishiraj Saha Roy, and Gerhard Weikum. 2023. Explainable Conversational Question Answering over Heterogeneous Sources via Iterative Graph Neural Networks. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR*. 643–653. <https://doi.org/10.1145/3539618.3591682>
- [7] DBLP release. 2022. <https://dblp.org/rdf/release/dblp-2022-06-01.nt.gz>
- [8] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, and et al. 2024. DeepSeek-V3 Technical Report. *CoRR* abs/2412.19437 (2024). arXiv:2412.19437 <https://arxiv.org/abs/2412.19437>
- [9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT)*. 4171–4186. <https://doi.org/10.18653/v1/n19-1423>
- [10] Dennis Diefenbach, Kamal Deep Singh, and Pierre Maret. 2018. WDAqua-core1: A Question Answering service for RDF Knowledge Bases. In *Companion Proceedings of the The Web Conference, (ACM)*. 1087–1091. <https://doi.org/10.1145/3184558.3191541>
- [11] YAGO downloads. 2022. <https://yago-knowledge.org/downloads/yago-4>
- [12] Mohnish Dubey, Debayan Banerjee, Debanjan Chaudhuri, and Jens Lehmann. 2018. EARL: Joint Entity and Relation Linking for Question Answering over Knowledge Graphs. In *Proceedings of the International Semantic Web Conference (ISWC)*, Vol. 11136. 108–126. [https://doi.org/10.1007/978-3-030-00671-6\\_7](https://doi.org/10.1007/978-3-030-00671-6_7)
- [13] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitan, Robert Oszuwa, and Jonathan Larson. 2024. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv preprint arXiv:2404.16130* (April 2024). <https://doi.org/10.48550/arXiv.2404.16130> [cs.CL] v2 revised 19 Feb 2025.
- [14] Sen Hu, Lei Zou, Jeffrey Yu, Haixun Wang, and Dongyan Zhao. 2018. Answering Natural Language Questions by Subgraph Matching over Knowledge Graphs. *IEEE Transactions on Knowledge and Data Engineering, (TKDE)* 30 (2018), 824–837. <https://doi.org/10.1109/TKDE.2017.2766634>
- [15] Xixin Hu, Yiheng Shu, Xiang Huang, and Yuzhong Qu. 2021. EDG-Based Question Decomposition for Complex Question Answering over Knowledge Bases. In *Proceedings of the International Semantic Web Conference, (ISWC)*. 128–145. [https://doi.org/10.1007/978-3-030-88361-4\\_8](https://doi.org/10.1007/978-3-030-88361-4_8)
- [16] IBM Research. 2024. Granite Foundation Models. <https://www.ibm.com/downloads/documents/us-en/10a99803c92fdb35>
- [17] Gautier Izacard and Edouard Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics: EACL*. 874–880. <https://doi.org/10.18653/V1/2021.EACL-MAIN.74>
- [18] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, and et al. 2023. Mistral 7B. *CoRR* abs/2310.06825 (2023). arXiv:2310.06825 <https://doi.org/10.48550/arXiv.2310.06825>
- [19] Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, and et al. 2025. Gemma 3 Technical Report. *CoRR* abs/2503.19786 (2025). arXiv:2503.19786 <https://doi.org/10.48550/arXiv.2503.19786>
- [20] Pavan Kapanipathi, Ibrahim Abdelaziz, Srinivas Ravishankar, Salim Roukos, and Alexander G. Gray et al. 2021. Leveraging Abstract Meaning Representation for Knowledge Base Question Answering. In *Findings of the Association for Computational Linguistics: (ACL/TJCNLP)*. 3884–3894. <https://doi.org/10.18653/v1/2021.findings-acl.339>
- [21] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 39–48.
- [22] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems (NeurIPS)*. [http://papers.nips.cc/paper\\_files/paper/2022/hash/8bb0d291acd4acf06f12099c16f326-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06f12099c16f326-Abstract-Conference.html)
- [23] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, and et al. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of Symposium on Operating Systems Principles, SOSP*. 611–626. <https://doi.org/10.1145/3600006.3613165>
- [24] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, and et al. 2015. DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web* 6 (2015), 167–195. <https://doi.org/10.3233/SW-140134>
- [25] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, and et al. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL*. 7871–7880. <https://doi.org/10.18653/v1/2020.ACL-MAIN.703>
- [26] Xue Li and Till Döhmen. 2024. Towards efficient data wrangling with llms using code generation. In *Proceedings of the Eighth Workshop on Data Management for End-to-End Machine Learning*. 62–66.
- [27] LMSYS Org. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90% ChatGPT Quality. <https://lmsys.org/blog/2023-03-30-vicuna/>
- [28] MAG records. 2022. <https://zenodo.org/record/4617285#YrNszNLMJhH>
- [29] Yohei Nakajima. 2023. Task-driven autonomous agent. <https://github.com/yoheinakajima/babyagi>
- [30] Reham Omar, Ishika Dhall, Panos Kalnis, and Essam Mansour. 2023. A Universal Question-Answering Platform for Knowledge Graphs. *Proceedings of the ACM on Management of Data* 1, 1 (2023), 57:1–57:25. <https://doi.org/10.1145/3588911>
- [31] Reham Omar, Omij Mangukiya, and Essam Mansour. 2025. Dialogue Benchmark Generation from Knowledge Graphs with Cost-Effective Retrieval-Augmented LLMs. *Proceedings of the ACM on Management of Data* 3, 1 (2025), 31:1–31:26. <https://doi.org/10.1145/3709681>
- [32] OpenAI. 2024. GPT-4o. (2024). <https://openai.com/index/hello-gpt-4o/>
- [33] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, and et al. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of machine learning research* 21 (2020), 140:1–140:67. <https://jmlr.org/papers/v21/20-074.html>
- [34] Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, and et al. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *CoRR* abs/2408.00118 (2024). arXiv:2408.00118 <https://doi.org/10.48550/arXiv.2408.00118>
- [35] Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, and et al. 2023. Code Llama: Open Foundation Models for Code. *CoRR* abs/2308.12950 (2023). arXiv:2308.12950 <https://doi.org/10.48550/arXiv.2308.12950>
- [36] Ahmad Sakor, Isaiah Mulang, Kuldeep Singh, Saeedeh Shekarpour, and Maria-Esther Vidal et al. 2019. Old is Gold: Linguistic Driven Approach for Entity and Relation Linking of Short Text. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT)*. 2336–2346. <https://doi.org/10.18653/v1/n19-1243>
- [37] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 3715–3734.
- [38] Liang Shi, Zhengju Tang, Nan Zhang, Xiaotong Zhang, and Zhi Yang. 2025. A Survey on Employing Large Language Models for Text-to-SQL Tasks. *ACM Comput. Surv.* (2025). <https://doi.org/10.1145/3737873>
- [39] Significant Gravitas. 2023. Auto-GPT: An Autonomous GPT-4 Experiment. <https://github.com/Significant-Gravitas/Auto-GPT>

- [40] Ion Stoica, Matei Zaharia, Joseph Gonzalez, Ken Goldberg, Koushik Sen, and et al. 2024. Specifications: The missing link to making the development of LLM systems an engineering discipline. *CoRR* abs/2412.05299 (2024). arXiv:2412.05299 <https://doi.org/10.48550/arXiv.2412.05299>
- [41] Thomas Tanon, Gerhard Weikum, and Fabian Suchanek. 2020. YAGO 4: A Reasonable Knowledge Base. In *Proceedings of the European Semantic Web Conference, (ESWC)*, Vol. 12123. 583–596. [https://doi.org/10.1007/978-3-030-49461-2\\_34](https://doi.org/10.1007/978-3-030-49461-2_34)
- [42] Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR* abs/2403.05530 (2024). arXiv:2403.05530 <https://doi.org/10.48550/arXiv.2403.05530>
- [43] Priyansh Trivedi, Gaurav Maheshwari, Mohnish Dubey, and Jens Lehmann. 2017. LC-QuAD: A Corpus for Complex Question Answering over Knowledge Graphs. In *Proceedings of the International Semantic Web Conference (ISWC)*, Vol. 10588. 210–218. [https://doi.org/10.1007/978-3-319-68204-4\\_22](https://doi.org/10.1007/978-3-319-68204-4_22)
- [44] Ricardo Usbeck, Ria Gusmita, Axel-Cyrille Ngomo, and Muhammad Saleem. 2018. 9th Challenge on Question Answering over Linked Data (QALD-9). In *Joint proceedings of the Workshop on Semantic Deep Learning (SemDeep-4) and NLIWoD4: Natural Language Interfaces for the Web of Data (NLIWOD-4) and 9th Question Answering over Linked Data challenge (QALD-9) co-located with International Semantic Web Conference (ISWC)*, Vol. 2241. 58–64. <http://ceur-ws.org/Vol-2241/paper-06.pdf>
- [45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, and Llion Jones et al. 2017. Attention is All you Need. In *Proceedings of the Advances in neural information processing systems (NeurIPS)*, 5998–6008. <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
- [46] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85. <https://doi.org/10.1145/2629489>
- [47] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science* 18, 6 (2024).
- [48] Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. 2024. Executable code actions elicit better llm agents. In *Forty-first International Conference on Machine Learning*.
- [49] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, and et al. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems (NeurIPS)*. [http://papers.nips.cc/paper\\_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html)
- [50] Tianbao Xie, Fan Zhou, Zhoujun Cheng, Peng Shi, Luoxuan Weng, and et al. 2023. OpenAgents: An Open Platform for Language Agents in the Wild. *CoRR* abs/2310.10634 (2023). arXiv:2310.10634 <https://arxiv.org/abs/2310.10634>
- [51] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, and et al. 2024. Qwen2.5 Technical Report. *CoRR* abs/2412.15115 (2024). arXiv:2412.15115 <https://doi.org/10.48550/arXiv.2412.15115>
- [52] Ke Yang, Yao Liu, Sapana Chaudhary, Rasool Fakoor, Pratik Chaudhari, George Karypis, and Huzefa Rangwala. 2024. Agentoccam: A simple yet strong baseline for llm-based web agents. *arXiv preprint arXiv:2410.13825* (2024).
- [53] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, and et al. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *The International Conference on Learning Representations, ICLR*. [https://openreview.net/forum?id=WE\\_vluYUL-X](https://openreview.net/forum?id=WE_vluYUL-X)
- [54] Lei Zou, Ruizhe Huang, Haixun Wang, Jeffrey Yu, and Wenqiang He et al. 2014. Natural language question answering over RDF: a graph data driven approach. In *Proceedings of the International Conference on Management of Data, (SIGMOD)*. 313–324. <https://doi.org/10.1145/2588555.2610525>