Schema Aware Conversational Data Collection

Lucas Emmanuel de Sousa Alves[1], **Claudio E. C. Campelo**[1]

[1]Systems and Computing Department – Federal University of Campina Grande (UFCG)
58429-900 – Campina Grande – PB – Brazil

lucas.alves@ccc.ufcg.edu.br, campelo@dsc.ufcg.edu.br

Abstract. *This paper proposes a conversational approach to structured data collection using an automated agent that addresses limitations of traditional form-based methods. Guided by a predefined schema of attributes, types, and constraints, the agent conducts dialogues to clarify ambiguities, complete missing information, and validate data in real time while minimizing the number of interactions. To support this, we developed a Python library enabling rapid creation of applications that proactively request user input. Using this library, we built a chatbot evaluated through user-based experiments in a culinary habits case study. Results show the approach effectively validates constraints, supports semantic inference, and guides users to provide complete and consistent data.*

1. Introduction

Structured data collection is a fundamental requirement in most computational systems. For this purpose, forms remain widely used because they provide strict control over fields and validations. Despite these advantages, they present significant limitations: they do not perform semantic inference, fail to provide real-time feedback when responses are inconsistent, and do not dynamically adapt to the context of the interaction.

A more natural alternative is dialogue driven by chatbots, in which a conversational agent progressively extracts information over multiple turns. Despite this potential, data collected through natural language must be mapped to structured schemas with clearly defined types, required fields, and constraints. However, existing approaches to structured data extraction from text focus primarily on the analysis of static documents or the processing of completed dialogues, including those using LLMs [Luo et al. 2025, Xu et al. 2024a, Zhang et al. 2022, Wei et al. 2024], and are not designed to actively guide a conversation toward the structured and validated collection of information.

Within dialogue systems, Dialogue State Tracking (DST) represents the most established approach to this problem and can be addressed through models trained to handle fixed slots in specific domains [Henderson et al. 2014], which requires the development of new architectures for each context. As a more flexible alternative, schema-guided approaches enable operation in generalized scenarios without the need for task-specific prior training [Rastogi et al. 2020]; however, they typically focus on tracking the conversational state while delegating constraint validation to later stages. Multi-agent architectures based on LLMs further expand dialogue management capabilities, but they are not specifically designed to ensure compliance between the collected information and a structured schema with typed constraints throughout the conversation.

This work proposes a conversational approach for active structured data collection, guided by a schema and based on a multi-agent architecture using LLMs. The approach proactively guides the dialogue, addresses semantic ambiguities, enforces constraints in real time, and automatically generates validated JSON output, thereby reducing manual mapping effort and promoting greater cross-domain generalization. To demonstrate its feasibility, the approach was implemented and evaluated through a case study involving 31 participants.

2. Related work

Several lines of research are relevant to the problem addressed in this work. In the context of zero-shot and multi-domain DST, [Tavares et al. 2023] formulate dialogue state tracking as a Question Answering task by converting each slot into a question and using minimal field representations to eliminate the need for domain-specific annotations. [Mehri and Eskénazi 2021] propose the Schema Attention Model (SAM), a graph-based architecture built on BERT-base that enables the system to determine subsequent actions in multi-domain tasks without task-specific prior training. [Zhao et al. 2022] demonstrate that using natural language descriptions of schema fields, rather than rigid structures, significantly improves model understanding and generalization, introducing D3ST, which employs an index-picking mechanism to generate slots and intents as output.

In the area of structured information extraction from text, [Luo et al. 2025] present a schema-guided knowledge extraction system based on a multi-agent architecture using LLMs, designed to identify and structure information from sources such as web pages, PDFs, and completed dialogues. [Xu et al. 2024b] propose CoTE (Chain-of-Thought Explanation), a method aimed at improving slot-filling accuracy through multi-step reasoning, thereby making the extraction process more transparent.

Although these studies advance the fields of DST and schema-guided extraction, none specifically focuses on actively guiding dialogue while continuously validating constraints throughout the interaction. [Luo et al. 2025], in particular, shares a multi-agent architecture and a schema-guided strategy with our work, but is designed for extracting pre-existing knowledge. As a result, a gap remains in approaches aimed at the active and validated collection of information directly from users during dialogue. Our approach addresses this gap by combining question generation grounded in LLM reasoning with a multi-agent architecture designed for dialogue-based data extraction, guided by a schema whose descriptions may include restrictive validations, while ensuring a structured output that is immediately ready for integration with external systems.

3. Methodology

The approach adopts a schema-guided strategy combined with a multi-agent architecture using LLMs, operating in a zero-shot setting without the need for task-specific prior training for each domain. The implementation leverages LangGraph[1] and LangChain[2] to orchestrate the agents, enabling the creation of iterative and conditional workflows that support the re-evaluation of information throughout the interaction.

[1]https://www.langchain.com/langgraph
[2]https://www.langchain.com/

3.1. Schema specification

The schema is one of the core elements of the approach, as it defines the rules that guide the entire extraction process. It is structured in JSON format and included in the prompt of all agents, ensuring that each agent accurately understands the context guiding the data collection process. Each field to be collected must follow the template shown in Figure 1, including a natural language description, the expected data type, and an indication of whether it is required. More detailed descriptions, including explicit constraints and examples, tend to help the agents comply with the defined rules, particularly in cases involving semantically ambiguous terms.

```
{
  "<field_name>": {
      "description": "<field_description>",
      "type": "<expected_data_type>",
      "required": <boolean>
  }
}
```

Figure 1. Structural template of a schema field

3.2. Multi-agent architecture

The approach adopts five specialized agents—Starter, Enquirer, Extractor, Checker, and Output—which follow a controlled workflow, as illustrated in Figure 2. The conversational state is maintained in a shared structure, called AgentState, which stores the message history, the most recent question, the data already extracted, the pending fields, and the final structured output throughout the interaction, allowing each agent to make decisions based on the full conversational context.

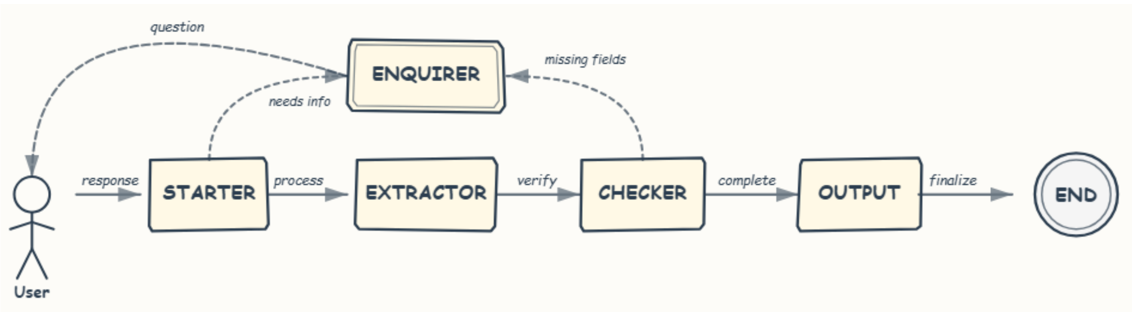


Figure 2. Solution architecture

The Starter serves as the entry point, classifying the user's intent and directing the workflow toward either information extraction or the collection of new information. When sufficient information is available, the Extractor is activated to process and normalize the data according to the schema; otherwise, the Enquirer takes over, formulating the next question in a natural and contextualized manner, with each prompt focused on a single pending field. After each extraction step, the Checker verifies whether any required fields remain uncollected, controlling the system's level of persistence until the Output formats the collected data into a standardized JSON structure with metadata and indicators of missing fields. Each agent is guided by complementary prompt engineering strategies, including role assignment, few-shot prompting, constraint-based prompting, and step-by-step reasoning instructions. The complete prompts are available in the public implementation repository[3].

3.3. Case study

To demonstrate the feasibility of the proposed approach, a case study was conducted on the collection of personal culinary preferences and habits, with the goal of evaluating the system's behavior in scenarios that go beyond direct and well-formed responses. The schema, presented in Table 1, consists of six fields with varying data types, requirement levels, and constraint complexity, introducing the following challenges:

1. **Explicit boundaries:** Fields impose minimum (M) and maximum (X) constraints, while the user may provide $M - Z$ or $X + Y$ items, where $Z, Y > 0$.
2. **Response flexibility:** Fields with different requirement levels require the agent to decide when to insist on a response or accept omission as valid.
3. **Specific data types:** Responses in natural language may require normalization to match the expected data types (`List`, `Number`, and `String`).
4. **Conditional field:** A field must be collected only if a specific schema condition is satisfied.
5. **Semantic inference:** Fields whose descriptive domain allows information to be implicitly inferred from the user's response.

Table 1. Excerpt from the schema used in the culinary preferences scenario.

Schema field specifications		
`preferred_cuisines`	**Type:** `List`	**Required:** Yes
Two or more cuisines that the user most frequently prefers (e.g., Italian, Japanese, Northeastern Brazilian).		
`favorite_locations`	**Type:** `List`	**Required:** Yes
Up to three places (restaurants, snack bars, or other establishments) where the user most enjoys eating or ordering food.		
`preferred_order_time`	**Type:** `String`	**Required:** No
Time of day (morning, afternoon, or evening). Collected only if weekly frequency ≥ 2. Allows semantic inference.		

The complete schema is available in the public implementation repository.

For example, when a user mentions only one dish, such as "pasta", the approach infers Italian cuisine and requests confirmation, thereby preserving the field's minimum constraint. When the user provides more items than allowed, the system intervenes by requesting a selection within the defined limit.

3.4. Experiment with volunteer participants

The evaluation was conducted with 31 participants through a web interface developed using Streamlit[4] and powered by the ChatGPT-4 model. This sample size aligns with

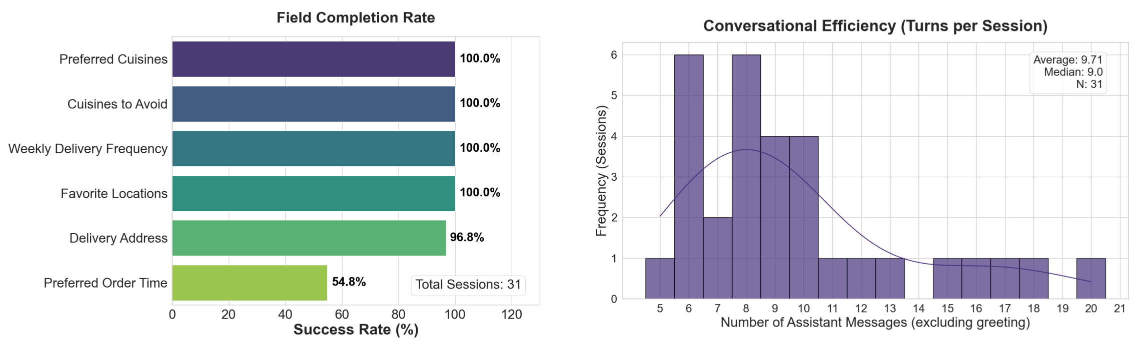
[3] https://github.com/ChatBoB-UFCG/ChatBoB

[4] https://streamlit.io

qualitative research literature indicating that behavioral patterns and data saturation are typically reached within this range [Guest et al. 2006]. Participants were instructed to challenge the dialogue rules in at least two interactions—for example, by providing a different number of items than requested, giving indirect responses, or refusing to answer certain questions—while responding normally in the remaining interactions, using realistic but not necessarily truthful answers. This design ensured that each session contained a natural mix of adversarial and cooperative turns, rather than exclusively adversarial behaviour. The evaluation was based on the following metrics: field completion rate, turns per session, constraint adherence rate, and semantic inference capability. The collected data are publicly available in the project repository.

4. Results and discussion

All sessions resulted in a structured JSON response without technical failures, strictly following the format defined by the Output agent and ensuring a fixed standard that enables the reuse of the output in automation workflows. Figure 3(a) presents the field completion rate. The fields preferred cuisines, cuisines to avoid, weekly delivery frequency, and favorite locations achieved a 100% completion rate. The delivery address field reached 96.8% compliance, with a single case in which the Starter agent incorrectly routed the workflow to the Enquirer agent instead of triggering extraction. The preferred order time field, being conditional, was completed in 54.8% of sessions—precisely in the cases where the weekly frequency condition of at least two orders was satisfied—while the remaining 45.2% were correctly classified as justified omissions.



(a) (b)

Figure 3. (a) Field completion rate; (b) distribution of the number of turns per session.

Figure 3(b) illustrates the number of turns per session. The approach achieved a median of 9 questions and reached up to 20 questions in a single dialogue to collect at most 6 fields, demonstrating that the system dynamically adapts its interaction strategy to ensure compliance with schema constraints without unnecessarily insisting on fields that have already been properly completed. The higher turn counts reflect deliberate user resistance as part of the experimental protocol, rather than redundant agent behavior

With respect to adherence to explicit constraints, shown in Figure 4, the preferred cuisines field achieved 100% compliance with the minimum requirement of two items, while the favorite places field maintained the maximum constraint of three items across all sessions. The cuisines to avoid field showed non-compliance in 19.4% of cases, corresponding to situations in which users reported having no dietary restrictions and the approach interpreted such responses as semantically valid, accepting an empty list—which violates the schema's minimum-item constraint. For the conditional preferred order time field, the rule was respected in 90.4% of sessions, with 9.7% of cases in which the question was asked incorrectly despite the weekly frequency being lower than 2.

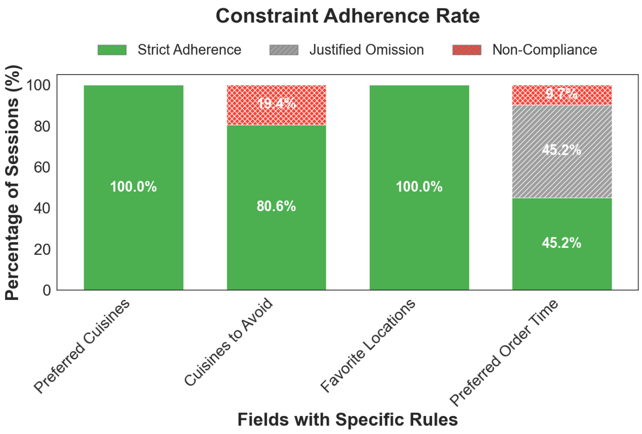


Figure 4. Constraint compliance for each field.

Regarding semantic inference, the approach correctly mapped implicit expressions to schema categories (e.g., "dinner" to "evening", "lunch" to "afternoon"). More detailed schema descriptions reduced divergent interpretations, while open-ended fields led to greater variation. Identified errors are addressable via prompt engineering or model selection and do not compromise the approach's overall robustness under adversarial conditions.

5. Conclusion and future work

The results demonstrate that the proposed approach is effective for the active collection of structured data through dialogue, showing robustness throughout the DST pipeline—from conversation management to the normalization and structuring of information—even in the presence of deliberate attempts to violate schema constraints. As future work, we intend to refine the prompt engineering strategies to better handle more complex scenarios, including the exploration of approaches such as Chain-of-Thought Explanation [Xu et al. 2024b]. Additionally, we plan to enhance the Extractor agent to perform simultaneous validation and extraction across multiple fields, preventing failures such as the one observed. Finally, we will conduct broader experiments across additional domains and a larger participant sample—since the current one limits statistical generalization—also investigating user engagement as a relevant variable not explicitly controlled in this study.

Acknowledgments and Declarations

Funding: This work was supported by the National Institute of Science and Technology in Responsible Artificial Intelligence for Computational Linguistics, Information Treatment, and Dissemination (INCT-TILDIAR), funded by CNPq, grant no. 408490/2024-1.

AI Statement: The Claude Sonnet 4.6 model was used exclusively for grammatical and stylistic review, with no impact on the intellectual content.

References

Guest, G., Bunce, A., and Johnson, L. (2006). How many interviews are enough? *Field Methods - FIELD METHOD*, 18:59–82.

Henderson, M., Thomson, B., and Williams, J. D. (2014). The second dialog state tracking challenge. In Georgila, K., Stone, M., Hastie, H., and Nenkova, A., editors, *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

Luo, Y., Ru, X., Liu, K., Yuan, L., Sun, M., Zhang, N., Liang, L., Zhang, Z., Zhou, J., Wei, L., Zheng, D., Wang, H., and Chen, H. (2025). Oneke: A dockerized schema-guided llm agent-based knowledge extraction system. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, page 2871–2874, New York, NY, USA. Association for Computing Machinery.

Mehri, S. and Eskénazi, M. (2021). Schema-guided paradigm for zero-shot dialog. *CoRR*, abs/2106.07056.

Rastogi, A., Zang, X., Sunkara, S., Gupta, R., and Khaitan, P. (2020). Schema-guided dialogue state tracking task at DSTC8. *CoRR*, abs/2002.01359.

Tavares, D., Semedo, D., Rudnicky, A., and Magalhaes, J. (2023). Learning to ask questions for zero-shot dialogue state tracking. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2118–2122, New York, NY, USA. Association for Computing Machinery.

Wei, X., Cui, X., Cheng, N., Wang, X., Zhang, X., Huang, S., Xie, P., Xu, J., Chen, Y., Zhang, M., Jiang, Y., and Han, W. (2024). Chatie: Zero-shot information extraction via chatting with chatgpt.

Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., Wu, X., Zheng, Y., Wang, Y., and Chen, E. (2024a). Large language models for generative information extraction: A survey.

Xu, L., Peng, N., Zhou, D., Ng, S.-K., and Fu, J. (2024b). Chain of thought explanation for dialogue state tracking.

Zhang, N., Xu, X., Tao, L., Yu, H., Ye, H., Qiao, S., Xie, X., Chen, X., Li, Z., and Li, L. (2022). DeepKE: A deep learning based knowledge extraction toolkit for knowledge base population. In Che, W. and Shutova, E., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 98–108, Abu Dhabi, UAE. Association for Computational Linguistics.

Zhao, J., Gupta, R., Cao, Y., Yu, D., Wang, M., Lee, H., Rastogi, A., Shafran, I., and Wu, Y. (2022). Description-driven task-oriented dialog modeling. *CoRR*, abs/2201.08904.