Journal of Computer Science & Technology

Indexed in: SCIE, EI, INSPEC, CBST, DBLP, etc.

Sponsored by: ICT, CAS & CCF

JCST Papers

Only for academic and non-commercial use

Thanks for reading!

JCST WeChat Subscription Account

JCST URL: https://jcst.ict.ac.cn

SPRINGER URL: https://www.springer.com/journal/11390

E-mail: jcst@ict.ac.cn

Online Submission: https://mc03.manuscriptcentral.com/jcst

Twitter: JCST_Journal

LinkedIn: Journal of Computer Science and Technology

Caffaro F, Rizzo G. Knowledge-enhanced conversational agents. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 39(3): 585-609 May 2024. DOI: 10.1007/s11390-024-2883-4

Knowledge-Enhanced Conversational Agents

Fabio Caffaro and Giuseppe Rizzo

LINKS Foundation, 10138 Turin, Italy

E-mail: fabio.caffaro@linksfoundation.com; giuseppe.rizzo@linksfoundation.com

Received October 3, 2022; accepted April 27, 2024.

Abstract Humanity has fantasized about artificial intelligence tools able to discuss with human beings fluently for decades. Numerous efforts have been proposed ranging from ELIZA to the modern vocal assistants. Despite the large interest in this research and innovation field, there is a lack of common understanding on the concept of conversational agents and general over expectations that hide the current limitations of existing solutions. This work proposes a literature review on the subject with a focus on the most promising type of conversational agents that are powered on top of knowledge bases and that can offer the ground knowledge to hold conversation autonomously on different topics. We describe a conceptual architecture to define the knowledge-enhanced conversational agents and investigate different domains of applications. We conclude this work by listing some promising research pathways for future work.

Keywords conversational agent, dialogue system, knowledge enhancing, artificial agent, intelligent conversation

1 Introduction

For a long time, humanity has fantasized about the possibility of creating intelligent artificial agents able to engage and interact with humans to assist them in performing some tasks or simply entertain them. The origin of this dream can be traced back much further than one can imagine. [1] reports how stories of creatures that were made, not born started to appear in ancient Greece around two thousand years ago. Greeks were also the first to develop the idea of ''automata'', i.e., self-operating machines designed to automatically follow a sequence of operations.

A cardinal aspect of an intelligent artificial agent is its ''capability to communicate''. With the explosion of the Digital Age, research moved towards the creation of software-based agents. These systems are generally referred to as conversational agents (CAs). In 1966, Dr. Joseph Weizenbaum from the Massachusetts Institute of Technology (MIT) presented ELIZA[2], the first CA able to interact with humans through a textual chat. Despite its simplicity, it had enormous success given its alleged capability to perform intelligent conversations. This led to the birth of the so-called ELIZA effect[3]. The incredible impact that this model had was probably due to the fact that it foreshadowed what the future of this technology could have brought. Indeed, it is not a case that in the following years the science fiction production was filled with many iconic artificial companions.

In more recent times, research efforts have moved towards conversational agents able to converse with humans with the purpose of ''assisting'' them in achieving a precise goal (e.g., booking a ticket for a flight). These models are commonly referred to as assistants. In 2011, 45 years after the launch of ELIZA, Apple's Siri was presented to the world as the first "intelligent assistant that helps you get things done just by asking"①.

Despite the rapid adoption of this technology, which led to the birth of many alternatives such as Google Assistant, Alexa and Cortana, internal analyses conducted by Amazon have uncovered concerns about user engagement, particularly in terms of retention, due to the users' rapid disinterest in the technology②. This swift lack of interest that plagues these technologies can be attributed to a phenomenon akin

Survey

①https://www.apple.com/nz/newsroom/2011/10/04Apple-Launches-iPhone-4S-iOS-5-iCloud, May 2024.

②https://www.bloomberg.com/news/articles/2021-12-22/amazon-s-voice-controlled-smart-speaker-alexa-can-t-hold-customer-interest-docs, May 2024.

to a new ELIZA effect, wherein users quickly become disillusioned due to high initial expectations that are not met.

Recently, large language models (LLMs) have emerged as the current new wave in the conversational technology, redefining the landscape of artificial intelligence and human-machine interactions. This wave has been propelled by advancements such as OpenAI's GPT series, including GPT-3[4], GPT-3.5[5], and GPT-4③, which showcase remarkable capabilities in generating impressive human-like text[4, 6]. The outstanding results obtained by these models, even at the early stages, had a major impact on public opinion too, to the extent that in 2020, GPT-3 was already elected AI "person" of the year by Forbes④. The advancements in LLMs have showcased remarkable performances across various tasks, yet these models can be defined as ''stochastic parrots''[7] due to their capacity to only mimic humans' conversational capabilities. The only knowledge they possess is embedded during the training phase and submerged within their vast network of weights.

Detailed analyses show how CAs tend to lose coherence and contradict themselves[4] particularly when they are interrogated on questions that involve socially important subjects such as morality and law, where they still have near-random accuracy[8]. Moreover, the adoption of LLMs in safety-critical domains has raised significant concerns, as these models often exhibit hallucination[9], generating unreliable responses in domain-specific or knowledge-intensive tasks[10], and particularly when queried on matters of critical importance[11].

A conversational interaction is a complex semantic activity, more difficult than "simple" language generation. [12] describes it as a process to create some meaning, where at turns two agents negotiate the meaning through the sharing of commonsense, personal, and social knowledge that has been acquired through experience or extensive study, aspects that language models may only partially capture from the statistical data in their training samples. For this reason, an important stream of research on CAs is currently focusing on developing new and efficient methodologies apt to enhance the agents' capabilities with external knowledge[13]. Additional knowledge can help agents have deeper semantic understanding of the conversation that cannot be inferred only by the conversation itself, and also provide more specific responses based on factual domain knowledge (e.g., FAQ documents) or offer customized services exploiting knowledge about user habits.

The objective of this work is to provide an exhaustive overview of the current developments in the technology supporting knowledge-enhanced conversational agents and their applications. This is done through a narrative literature review that addresses four general research questions (RQs).

The remaining of this work is structured as follows: Section 2 provides a general overview of the task, Section 3 goes into the implementation details of these models, Section 4 investigates different domains of application, Section 5 explores possible path for future improvements, and Section 6 provides a final discussion about the topic.

2 Task Definition (RQ1)

Traditional conversational AI has predominantly focused on natural language processing, dialogue management, and user intent recognition. While these systems have exhibited impressive capabilities, they often rely on pre-defined responses and may struggle with nuanced understanding and context-awareness. On the other hand, recent LLMs have demonstrated impressive results on different tasks, but are susceptible to hallucination, particularly in high-risk knowledge domains[11]. Knowledge-enhanced conversational agents (KCAs), through dynamic exploitation of various knowledge sources, offer the promise of providing contextually relevant responses and enhancing reliability, while also contributing to the explainability in conversational interactions[14]. This section will provide an analysis of the characteristics of both CA and KCA.

2.1 Conversational Agents

A conversational agent (CA), often referred to as a chatbot or virtual assistant[15], is a software or artificial intelligence system designed to engage in conversational interactions with users[16]. These agents facilitate human-computer interactions through text or voice-based communication.

③https://openai.com/index/gpt-4-research, May 2024.

④https://www.forbes.com/sites/kenrickcai/2021/01/04/forbes-ai-awards-2020-meet-gpt-3-the-computerprogram-that-can-write-an-op-ed/?sh=18aa83d693a7, May 2024.

2.1.1 Characteristics of CAs

CAs exhibit a spectrum of characteristics that distinguish them based on their design and functionality. In particular, they are characterized considering mainly four different criteria (Fig.1).

The first and most divisive criterion of distinction is the goal of conversation. This is the final scope for which the agent is programmed. In this sense, the agents can be distinguished as two types.

- *Task-Oriented*: designed to perform conversations with users within the scope to assist them to complete certain tasks (e.g., booking a flight ticket or a hotel room)[17].
- *Non-Task-Oriented*: whose goal is to engage in a chit-chat conversation with users, without any specific purpose, except that of maximizing their interest and engagement[18].

The second criterion of distinction is based on the domain on which the agent operates. Here we can distinguish three types of agents:

- *Single-Domain*: designed to operate on a single specific domain (e.g., flights);
- *Multi-Domain*: designed to operate on a finite set of domains (e.g., flights, hotels, and restaurants);
- *Open-Domain*: untied from any specific domain.

The third criterion regards the modality of the conversation, which is the channel adopted to convey the conversational interaction. Conventionally, this is related to natural language as text or voice. However, other modalities may also be considered, for example, videos and images, touch input, gestures. Nowadays, modern KCAs usually grant greater flexibility to the users allowing more than a single modality. These agents are referred to as multi-modal.

The last criterion is related to the conversation memory of the agent. This represents the dialogue history retained by the model during a single session and it can be considered as a sort of short-term memory of the agent that allows it to be more aware of the conversation that is happening. According to this criterion, there are two types of agents:

- *Single-Turn.* These agents do not keep any information from one turn to another. Each turn behaves like an independent act of the dialogue involving a single user question and the following system response. This type is called one-shot exchange.
- *Multi-Turn.* These agents at each turn are aware of a window composed by the previous n turns. This conversation memory is used for example to solve anaphoric references[5].

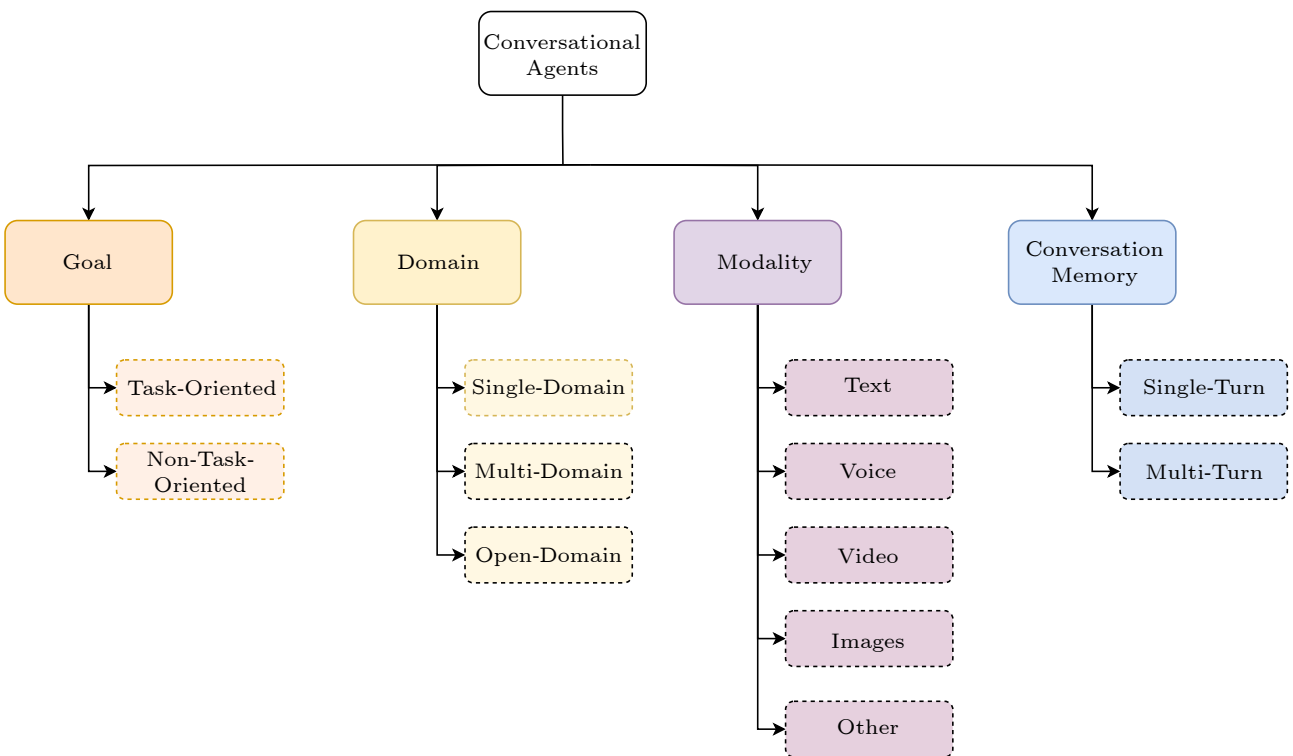

Fig.1. Four main properties that characterize a CA.

[5]An anaphoric reference is a word or phrase that refers to something mentioned previously.

2.1.2 Types of CAs

Conversational Agent is a general term that encloses a wide set of different agents. Chatbot[19] and Intelligent Personal Assistant[20], for instance, are often used as synonyms for CA, even if they refer to substantially different agents. Distinctly from "Conversational Agent" that can work as a general term, these are instances of a CA that explicitly specify the type of agent and the type of conversation. For example, the term chatbot refers to a CA, where the agent is a bot, and the conversation happens through a chat. Table 1 shows some examples of the most diffused CAs highlighting the type of conversation and the type of agent.

Based on the characteristics explained in Subsection 2.1.1, it is possible to identify three main categories of CAs that can be found on the literature: chatbots, conversational interfaces, and assistants.

Table 1. Popular Types of CAs Distinguished Based on the Type of Conversation and Agent

Conversation	Agent	Reference
Chat	Bot	Llama-3⑥
		Gemini[21]
		GPT-4⑦
		ChatGPT[5]
		PLATO2[22]
		Meena[23]
Social	Bot	Alexa Prize[16]
		ALANA[24]
		[25]
		Twitter Darpa challenge[26]
Q&A	Bot	BlenderBot[27]
		GraftNet[28]
Voice	(User) interface	[29]
		[30]
Conversational	(User) interface	StockBabble[31]
		[32]
		[33]
Personal	(Digital) assistant	Siri⑧
		Amazon Alexa⑨
		[34]
		PAL3[35]
Intelligent	(Virtual) assistant	Google Assistant⑩
		[20]
		[36]

Chatbots are conversational agents that can communicate with users through text or voice (also called voicebots)[37], and their primary focus is often to provide information, perform tasks, or engage in basic conversations. Typically, they provide non-task-oriented and open-domain conversation with a multi-turn conversation memory. The primary strength of this kind of models lies in their ability to generate contextually relevant responses in open-ended conversations, making them a valuable tool for natural language understanding and generation across various domains and topics. Recent examples of chatbots include widely-used LLM-based models like OpenAI's ChatGPT[5] and GPT-4⑦, Meta's LLama-3⑥, and Google's Gemini[21].

Conversational interfaces encompass a broader category of conversational agents that serve as the means of interaction between users and computer systems. These interfaces can be voice-based[29] or text-based[32] (also referred to as conversational user interfaces) and are integral in enabling natural language communication with technology providing a valid alternative to the classical graphical user interface (GUI). These agents provide a conversational interface towards a specific resource, service or application[33]. These agents are usually task-oriented and single-domain since they are tailored to the specific resource's needs. In addition, to reduce the complexity of these models, often the dialogue flow is designed a priori providing a set of predefined answers to the users. Commonly, these models are multi-turn since they need to have a memory of more than just a single turn to deal with complex queries. A classical example is Voice User Interface that is vastly adopted to provide automated customer support over the phone[37].

Assistants represent conversational agents designed to provide comprehensive support to users in various aspects of their daily lives. These models provide interactions usually in a multi-modal manner, incorporating text, voice, touch input, and multimedia elements. They can be personal assistants that offer assistance to individuals or digital assistants embedded in software or devices⑩. A distinctive trait, is their inclination toward being person-oriented. Ideally, they should possess and recall knowledge of the user's preferences, habits, and interests, aiming to offer

⑥https://ai.meta.com/blog/meta-llama-3, May 2024.
⑦https://openai.com/index/gpt-4-research, May 2024.
⑧https://www.apple.com/siri, May 2024.
⑨https://alexa.amazon.com, May 2024.
⑩https://assistant.google.com, May 2024.

personalized experience. However, it is important to note that their ability to achieve full personalization is currently limited. Instead, their interactions typically involve single-turn exchanges aimed at activating one of many available services. For example, with digital assistants like Alexa, the constrained nature of interactions, where each prompt typically begins with "Alexa...", naturally limits the dialogue to single-turn exchanges. Hence, their primary function reduces to discern the user's intent and trigger the corresponding action, rather than engaging in extended multi-turn conversations.

Some assistants can be dynamically extended to new services. However, they normally offer only the activation of the service. The development of the conversational interface within a service is left up to the owners of the resource providing a development kit such as Amazon ASK[11] or Google Action SDK[12], which can be used to design more complex multi-turn domain-specific conversations in order to handle a variety of queries.

In this context, an emerging and more flexible solution is based on the so-called function calling[13], a feature that allows open-domain models like GPT-4 to interface with external APIs, transforming them into hybrid assistants able to interact with external tools and services and performing a wide range of tasks, while offering a smooth conversational interface. This integration extends the basic model's capabilities beyond simple text generation, allowing it to perform actions like sending emails or fetching weather updates.

2.2 Knowledge-Enhanced Conversational Agents

Knowledge is a fundamental element to establish effective conversational interactions. For instance, considering chit-chat conversations, socially shared commonsense knowledge represents the background information that people use during conversations[38, 39]. Regarding task-oriented models instead, the additional knowledge can help in designing more robust models able to offer a frictionless conversational interaction with out-of-scope requests[40]. For these reasons, in recent times the research focused on designing CAs able to leverage additional knowledge to provide more specific responses.

A knowledge-enhanced conversational agent (KCA) represents an advanced category of conversational AI. It is characterized by its ability to access and integrate information from external knowledge sources or databases in real-time during conversations with users. Unlike rule-based or task-oriented agents, knowledge-enhanced CAs can provide contextually relevant and informative responses to a wide array of user queries, even on complex topics.

In its most simple form, a KCA can be thought of as a modular framework whose general architecture is composed of three main blocks (Fig.2): a dialogue system that deals with the conversational part of the agent, a memory system that manages its knowledge resources, and a knowledge-enhancing interface that harmonizes the information flow between the two preceding modules.

2.2.1 *Knowledge-Enhancing Problem*

A natural conversational interaction can be thought of as an ordered sequence of utterances $U = (u_1, u_2, \ldots, u_n)$. In the context of conversational AI, it is commonly assumed that the conversation happens in a turn-taking fashion, in which a user and an agent alternate their utterances. n indicates the number of turns of the interaction. Each utterance corresponds to an ordered sequence of tokens $u_i = (x_1, \ldots, x_l)$. The dialogue context $U_{t,\,w} = (u_{t-w+1}, \ldots, u_{t-1}, u_t)$ at time t and with window size w is defined as the ordered set of the last w utterances, ending with an utterance u_t belonging to the user.

Given in input a dialogue context $U_{t,\,w}$, a KCA aims at generating an appropriate response u_{t+1} grounded on a set of relevant knowledge documents $\tilde{K}$.

3 Implementation (RQ2)

In this section, we delve into the practical aspects of implementing a KCA. We explore three critical facets of implementation: the Dialogue System, responsible for managing the conversational flow; the Memory System, encompassing structured and unstructured knowledge bases; and the Knowledge-Enhancing Interface, bridging the gap between the CA and external knowledge resources. By dissecting these

[11] https://developer.amazon.com/en-US/alexa/alexa-skills-kit, May 2024.
[12] https://developers.google.com/assistant/df-asdk/actions-sdk, May 2024.
[13] https://openai.com/index/function-calling-and-other-api-updates, May 2024.

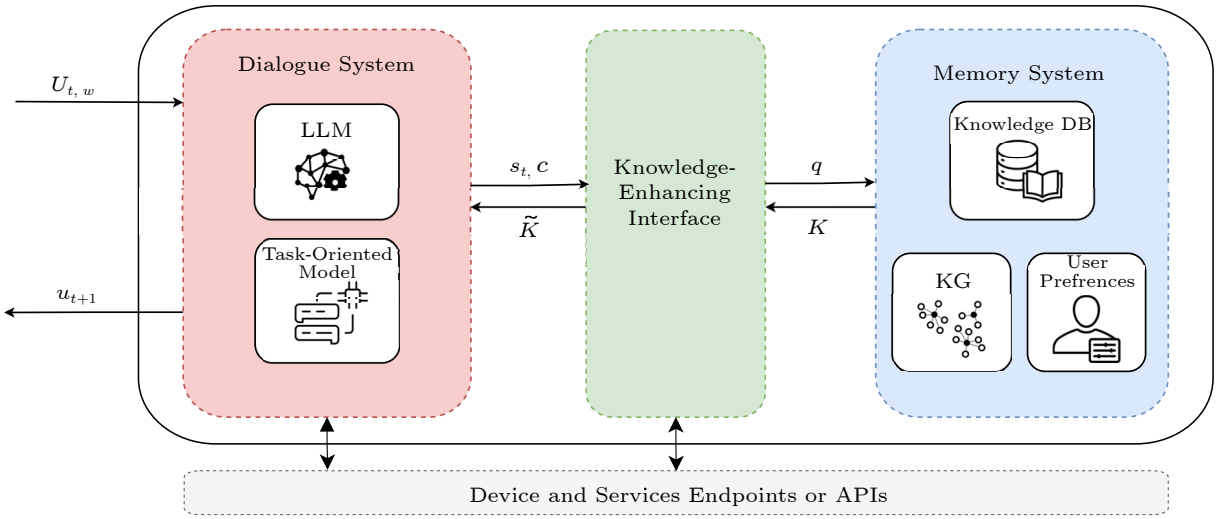

Fig.2. General architecture of a KCA. The dialogue system receives and preprocesses the dialogue context before passing it to the interface, which formulates a query to the memory systems. The interface filters and selects relevant knowledge, providing a seamless interaction between conversational agents and knowledge sources.

components, we aim to provide insights into the construction and integration of knowledge-enhanced CAs.

3.1 Dialogue System

The dialogue system (DS) is the module that deals with the conversational functionalities of the agent. This system has to manage the conversational input and prepare an adequate output response. In the case of a multi-modal agent, it has also to deal with the issues related to the combination of different modalities. This module is the core of the agent that mainly characterizes it. There exist two main architectures (see Fig.3) adopted to implement DSs: the pipeline architecture, and the end-to-end architecture.

3.1.1 *Pipeline Architecture*

In the pipeline architecture (see Fig.3(a)), the dialogue system consists of a modular structure in which the input information is transformed and passed sequentially through each module. This architectural approach is mostly common in task-oriented models, including conversational interfaces and assistants due to its fine-grained control over the conversational flow, enabling efficient handling of user queries and task-oriented interactions. The most diffused architecture is composed of three main modules named: lan-

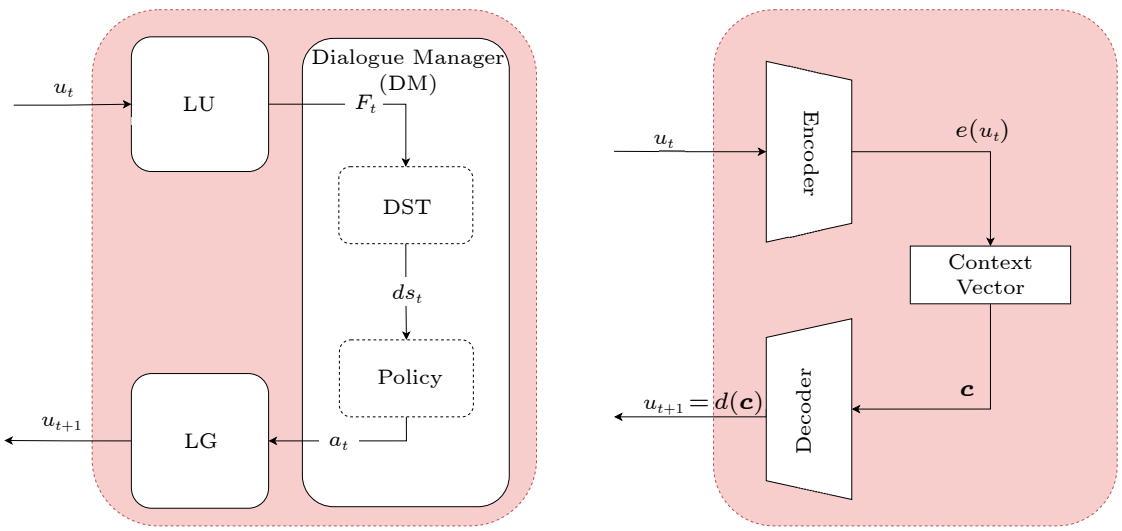

Fig.3. Two main architectures adopted to implement a dialogue system. (a) Pipeline architecture. (b) End-to-end architecture.

guage understanding (LU), dialogue manager (DM), and language generation (LG).

The LU module serves as the initial stage of the pipeline. It parses and interprets the user's messages, extracting essential information such as intents, entities, and context. LU's primary role is to comprehend what the user is saying or asking.

Given an input utterance $u_t = (x_1, x_2, \ldots, x_n)$, the LU module extracts a semantic representation[17] called semantic frame:

$$F_t = \arg\max_F P(F|u) = \langle d_u, i_u, \mathrm{s}_u \rangle.$$

Each semantic frame is typically composed by three parts of information: the domain of the utterance d_u, the intent i_u of the user, and a set of slots $s_u = \{s_1, \ldots, s_n\}$, each one of the type $s_i = \langle name, value \rangle$. Nowadays, the standard method adopted is based on large pre-trained Transformer models used to jointly perform intent-classification and slot-filling[41–43].

The DM module acts as the decision-maker within the pipeline architecture. It processes the output from LU and determines the dialogue act a_t, which is the next system's action based on the input semantic frame F_t. Generally, the DM module, is composed of two fundamental submodules.

- The dialogue state tracking (DST) submodule keeps track of information from the dialogue that is relevant for the system to choose the next action. This information is defined as the dialogue state ds_t and contains the abstract representations of the previous utterances of the conversation. In the case of task-oriented agents, it contains also the task record[37] that keeps track of the slots that have been already filled and the ones that are still needed in order to make a request to a service API. Recent approaches can be distinguished into ontology-based and ontology-free methods[44]. The former ones select a value from a candidate-value list for each target slot, while the latter ones offer more flexibility extracting the values as spans from dialogue contexts, predicting the start and end tokens of the value in the utterance[45, 46].
- The policy learning submodule generates the next system's action a_t based on the dialogue state[17]. In rule-based systems, these decisions are pre-scripted, with choices based on factors such as the confidence levels associated with the user's input[37]. Alternative approaches are based on the idea that a dialogue can be treated as a Markov decision process (MDP) where at each turn the agent moves from a (dialogue) state ds_t to a new state ds_{t+1} taking an action a_t. This task can be addressed using reinforcement learning (RL) techniques. MDP techniques require that the state of the agent is completely determinable and this is typically not the case for conversational interactions. Partially observable Markov decision processes (POMDP) tackle this problem considering a probability distribution over the possible current state in order to account for this uncertainty[47].

The LG module is responsible for crafting responses to be presented to the user. It takes the dialogue act output from the DM and generates human-readable and contextually relevant responses, which are then delivered back to the user. To improve the user experience, this output must be[40]:

- appropriate: in the sense that the response must be natural and informative according the input dialogue context U_t;
- accurate: in the sense that the semantics of the dialogue act must be fully conveyed.

In commercial systems, LG is often a fairly trivial task that involves the use of either canned text or pre-defined responses templates in combination with the substitution mechanism for the specific entities[37]. One of the main limitations of traditional template-based models is that they typically rely on heavily annotated data, which makes infeasible the generalization of the model to new domains. [48] proposes SC-GPT, a GPT model pre-trained on a large set of annotated corpus to acquire the controllable generation ability and then fine-tuned with only a few domain-specific labels to adapt to new domains.

3.1.2 End-to-End Architecture

One of the main limitations of pipeline models is their low capability of generalization, due to the fact that they are strictly constrained to the domains on which they are designed. For instance, RL-based systems require extensive handcrafting and design of the state and action space[37]. Furthermore, the presence of separate modules makes it difficult to trace which module is the cause for a failure in an interaction. This is referred to as the credit assignment problem[49]. In addition, the amelioration of a module does not necessarily translate into the improvements for the whole system[50].

To try to mitigate these problems, end-to-end models were proposed. In general, their aim is to gen-

erate a variable length output sequence $Y = (y_1, \ldots, y_m)$ conditioned on a variable length input sequence $X = (x_1, \ldots, x_n)$ learning directly the conditional distribution over Y given X:

$$P(Y \mid X) = P(y_1, \ldots, y_m \mid x_1, \ldots, x_n) = \prod_{t=1}^{m} P(y_t \mid x_1, \ldots, x_n, y_1, \ldots, y_{t-1}) .$$

The end-to-end architecture is a prominent approach in dialogue systems, especially suited for open-domain chit-chat conversations where the dialogue flow is less structured and more exploratory.

Classical end-to-end models work adopting the traditional encoder-decoder structure[51]. The encoder maps the input sequence $\boldsymbol{X} = (x_1, \ldots, x_n)$ into a latent representation $\boldsymbol{c}$ called context vector:

$$e : \boldsymbol{X} \longmapsto \boldsymbol{c}.$$

Transformer-based encoders such as ELECTRA[52], MPNet[53], and SimCSE[54], exploit the attention mechanism to compute global dependencies among all the input tokens.

Starting from a context vector the decoder produces the output sequence $\boldsymbol{Y}$. Transformer decoder models, are commonly trained according to a standard autoregressive objective (e.g., predicting the next word). LLMs, such as GPT-3[4] and PaLM[55], have played a pivotal role in advancing the capabilities of end-to-end DSs. These models, trained on massive corpora of text data, exhibit the ability to generate contextually relevant responses in open-ended conversations, making them well-suited for open-domain conversational applications.

Although LLMs can function as implicit knowledge bases[56], they are not reliable, particularly in high-risk fields such as science and information dissemination. A recent example highlighting this concern is Galactica[57], an LLM developed by Meta. Galactica was designed to store, combine, and reason about scientific knowledge, presenting itself as a potential solution to the challenge of information overload in scientific research. However, despite its impressive performance on various scientific tasks, Galactica's reliability was called into question. After just two days of public access, it became evident that the model was producing fake news and spreading misinformation. This incident underscores the inherent risks associated with relying solely on LLMs for critical tasks, especially in areas where accuracy and trustworthiness are paramount.

3.2 Memory System

The memory system $\mathcal{M}$ serves as the central repository for knowledge in knowledge-enhanced conversational agents. This is the place where knowledge sources, also called knowledge bases (KBs), are stored, organized, and accessed. This section explores the pivotal role of the memory system, its design considerations, and the types of knowledge bases commonly integrated into this module.

Different types of knowledge can be stored. [58] makes a distinction between internal and external sources of knowledge with respect to the conversation. The former ones are inferred from the input and can include for example keywords and linguistic features. The latter ones are provided from outside sources as knowledge bases. [59] differentiates between generic and domain-dependent knowledge. While the first offers additional information from any domain (e.g., information about user interaction like in [60]), the second is referred to a specific domain or set of domains[61].

Each knowledge source can be considered as a set of knowledge snippets that represent the atomic knowledge information of which the knowledge source is composed. Based on the form of the knowledge snippets, it is possible to distinguish between unstructured and structured knowledge sources.

3.2.1 *Structured Knowledge*

Structured knowledge sources are constructed by performing a map from the raw text to a semantic representation of the knowledge that then is stored in a specialized architecture. These sources include structured databases, knowledge graphs, and ontologies that organize information in a systematic and structured format. Knowledge graphs (KGs), in particular, have gained prominence in recent research, representing knowledge as interconnected entities and relationships. KGs store knowledge facts, each described as a triple of the form of the type $(entity, relation, entity)$[62]. These facts are organized into a directed graph structure, $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{F}\}$, where $\mathcal{E}$ is the set of entities that represent the nodes of the graph, $\mathcal{R}$ is the set of possible relations between two entities, and $\mathcal{F}$ is the set of knowledge facts (e_h, r, e_t), with $e_h, e_t \in \mathcal{E}$ and $r \in \mathcal{R}$ that represent the edges of the graph. Fig.4 displays an example of a KG.

KGs have been widely adopted to enhance the agent's responses, in non-task-oriented settings. Indeed, a KG is an ideal approach for combining com-

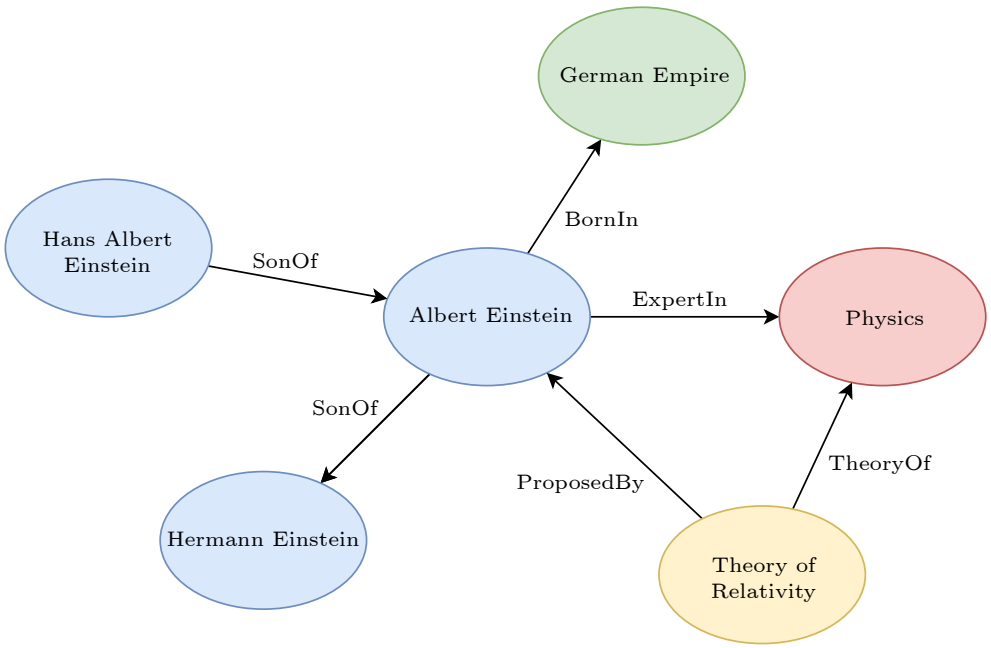

Fig.4. Simple example of a KG[62].

monsense knowledge into response generation[63]. [64] proposes ConceptFlow, a model representing the conversation flow as traverses in the commonsense relations contained in a KG. The traverse of the graph is guided by graph attentions, moving towards more meaningful directions in order to generate more informative responses.

However, structured knowledge sources are particularly advantageous for tasks that require reliability and interpretability. Typically in this case KGs are exploited to store domain-specific ontologies, which the agent can use to gain a deeper understanding of the input and generate more specific outputs. For instance, [65] uses a KG to incorporate domain-specific knowledge into an end-to-end task-oriented KCA. An intriguing and recent area of research focuses on the development of techniques apt to enhance LLMs performance by integrating external knowledge sources. [14] conducts a comprehensive study on enhancing LLMs with KGs to improve their factual reasoning abilities and performance in generating knowledge-grounded content. This research has spurred the development of knowledge graph enhanced large language models (KGLLMs) as a promising avenue for advancing the capabilities of language models.

3.2.2 *Unstructured Knowledge*

Unstructured knowledge sources represent the simplest form of knowledge, in which the knowledge snippets are typically raw-text documents. Although structured knowledge bases are preferable, unstructured knowledge represents the vast majority of the knowledge that can be mined from online resources such as Wikipedia for general knowledge, or Goodreads and Foursquare for domain-specific one[66]. Memory networks, can be leveraged to enhance KCAs with unstructured knowledge. Memory networks are recurrent models adopted to efficiently store and retrieve unstructured knowledge snippets $\{k_i\}$ in the form of embedded memory vectors $\{\boldsymbol{m}_i\}$. They were firstly introduced by [67] in the context of question answering (Q&A) as a long-term memory that can be read and written, and that can effectively act as a dynamic knowledge base. The core component of a memory network is the memory module m that is an array of memories $\boldsymbol{m}_i$ (e.g., an array of vectors or an array of strings that represent the knowledge).

The original memory network was not easy to train via backpropagation and required supervision at each layer of the network. [68] extends this work into an end-to-end model (see Fig.5) that requires significantly less supervision during training, making it more generally applicable in realistic settings. This memory network operates in three steps.

1) *Input Memory Representation.* The set of knowledge snippets $\{k_i\}$ is converted into memory vectors $\{\boldsymbol{m}_i = A(k_i)\}$ using a representation model A. The input utterance u_t is also converted into a query vector $\boldsymbol{q}_t = B(u_t)$ with a second representation model

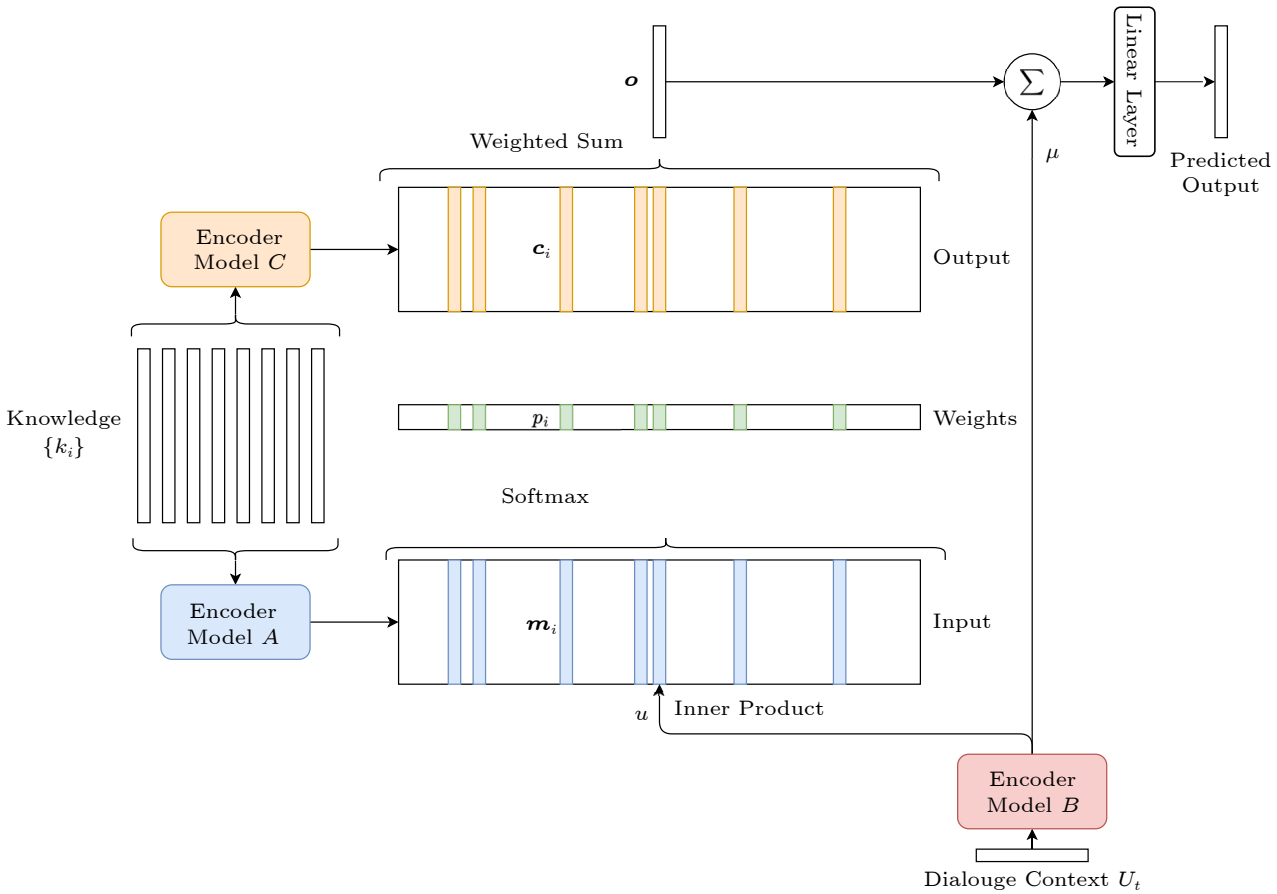

Fig.5. Basic architecture of the end-to-end memory network[68].

B. The match between each memory m_i and the query $\boldsymbol{q}_t$ are then computed as:

$$p_i = \text{Softmax}(\boldsymbol{q}_t^{\mathrm{T}} m_i).$$

2) *Output Memory Representation*. For each input k_i a corresponding output vector $\boldsymbol{c}_i = C(k_i)$ is computed with a third representation model C. The response vector from the memory is then computed as:

$$\boldsymbol{o} = \sum_i p_i \boldsymbol{c}_i.$$

3) *Final Selection*. The sum of the output vector $\boldsymbol{o}$ and the query vector $\boldsymbol{q}$ is passed through a final weight matrix W to compute the final scores:

$$w = \text{Softmax}(\boldsymbol{W}(o + q)),$$

which are then used to retrieve the final predicted answer $\hat{\boldsymbol{a}}$.

This model became the commonly accepted memory network, and is used as base for different models such as the famous key-value memory network[69].

Memory networks are a classic method adopted to deal with question-answering tasks. Firstly they were applied in the context of chit-chat conversations[66, 70] and then extended to task-oriented ones[71, 72]. Recently, memory networks have been adopted in the context of emotion recognition for conversational agents, as they can effectively learn and store long-term dependencies in conversations. DIMMN[73] uses a multi-view memory network to fuse multimodal data from different perspectives, which solves the emotional inconsistency of the model from different perspectives and enhances the robustness and accuracy of the algorithm in real scenarios.

3.3 Knowledge-Enhancing Interface

Regarding the implementation of the knowledge-enhancing interface, there are two main approaches. The first one consists in defining implicit interfaces heavily depending on the type of the dialogue system and knowledge adopted, the other consists in designing an explicit and independent interface composed of specialized blocks.

3.3.1 Implicit Interfaces

To perform the knowledge-enhancing process, traditional approaches adopt specialized architectures heavily depending on the type of the dialogue system and the type of knowledge source[58]. For instance, at the DM stage of the pipeline architecture, the agent works in a discrete setting, in which key information about the conversation (e.g., the domain of the request) are explicitly contained in the semantic frame and the dialogue state. This information can be used to construct an explicit query to a structured knowledge base in the memory system (see Fig.6(a)).

Contrarily, end-to-end dialogue systems work in continuous environments. The semantic representation of user input is a context vector in a latent space. In this case, the query to the memory system must be constructed directly from the context vector. A straightforward approach would be adopting context vector $\boldsymbol{c}$ as query vector $\boldsymbol{q}_t$ for an end-to-end memory network (see Fig.6(b)). An alternative approach involves generating intermediate latent representation[74]. For example, an attention mechanism can be exploited to retrieve a knowledge-aware context vector that will be used in the decoder to condition output sequence generation. Examples of these knowledge-attention mechanisms have been used in combination with knowledge graphs[75] for commonsense[64] and exploiting heterogeneous sources[76].

3.3.2 Explicit Interfaces

An alternative approach consists in designing an explicit interface composed of specific modules. In the context of LLMs, this approach is often referred to as Retrieval-Augmented Generation (RAG)[13]. RAG systems incorporate explicit knowledge retrieval mechanisms, allowing the model to access external knowledge sources to enhance its understanding and generation capabilities. These modules enable the model to retrieve relevant information from knowledge graphs, databases, or other structured data sources, which can then be seamlessly integrated into the generation process. Following the work of [40], the knowledge-enhancing task can be decomposed into three general subtasks, which can be associated to three corresponding modules (see Fig.7):

1) detection of turns that needs to access to the memory system $\mathcal{M}$ of the agent;

2) selection from the memory system of the contextually relevant knowledge;

3) generation of an appropriate response conditioned on the input and the knowledge selected.

1) *Detection*

The detection system determines whether the model's output should rely only on previous historical session content or need access to some additional knowledge. For each turn t, given a dialogue context U_t, the detection system detects if the current input utterance u_t requires additional knowledge from the memory system $\mathcal{M}$ in order to be addressed. The detection task can be treated as a binary classification problem on the dialogue context:

$$f_D(U_{t,w}) = \begin{cases} 1, & \text{if } U_{t,w} \text{ requires additional knowledge,} \\ 0, & \text{else.} \end{cases}$$

The basic approach to addressing this task consists in encoding $U_{t,w}$ into a context vector, which is then used as a feature vector for a traditional classification method. However, as explained by [77] and [78], this simple approach tends to easily overfit on frequent keywords or semantic patterns present in the domain contexts from the training set. Consequentially, the final detection system will suffer from poor generalization capabilities on unseen domains, having low performances on zero or few-shot settings. In this regard, [79] proposes a method called Representation Learning and Density Estimation (REDE) able to quickly learn a knowledge-seeking turn detector with just few out-of-domain samples.

An alternative approach to this problem consists of conditioning the classification problem on the dialogue context and some additional internal or external knowledge k_a:

$$f_D(U_{t,w}, k_a) = \begin{cases} 1, & \text{if } U_{t,w} \text{ requires additional knowledge,} \\ 0, & \text{else.} \end{cases}$$

[80] proposes a knowledge-aware ELECTRA model that exploits internal knowledge from the dialogue context itself, performing an entity tracking to dialogue context U_t. [81] introduces a novel method Self-Knowledge Guided Retrieval Augmentation (SKR), aimed at improving LLMs by enabling them to recognize their own knowledge boundaries and selectively utilize external information when necessary. This approach has been shown to outperform fully retrieval-based methods, marking a significant advancement in the field of LLMs and the application in question-answering tasks, suggesting that eliciting self-knowledge in LLMs can lead to more accurate responses.

Although the detection system is still under-explored, the importance of this module is that it al-

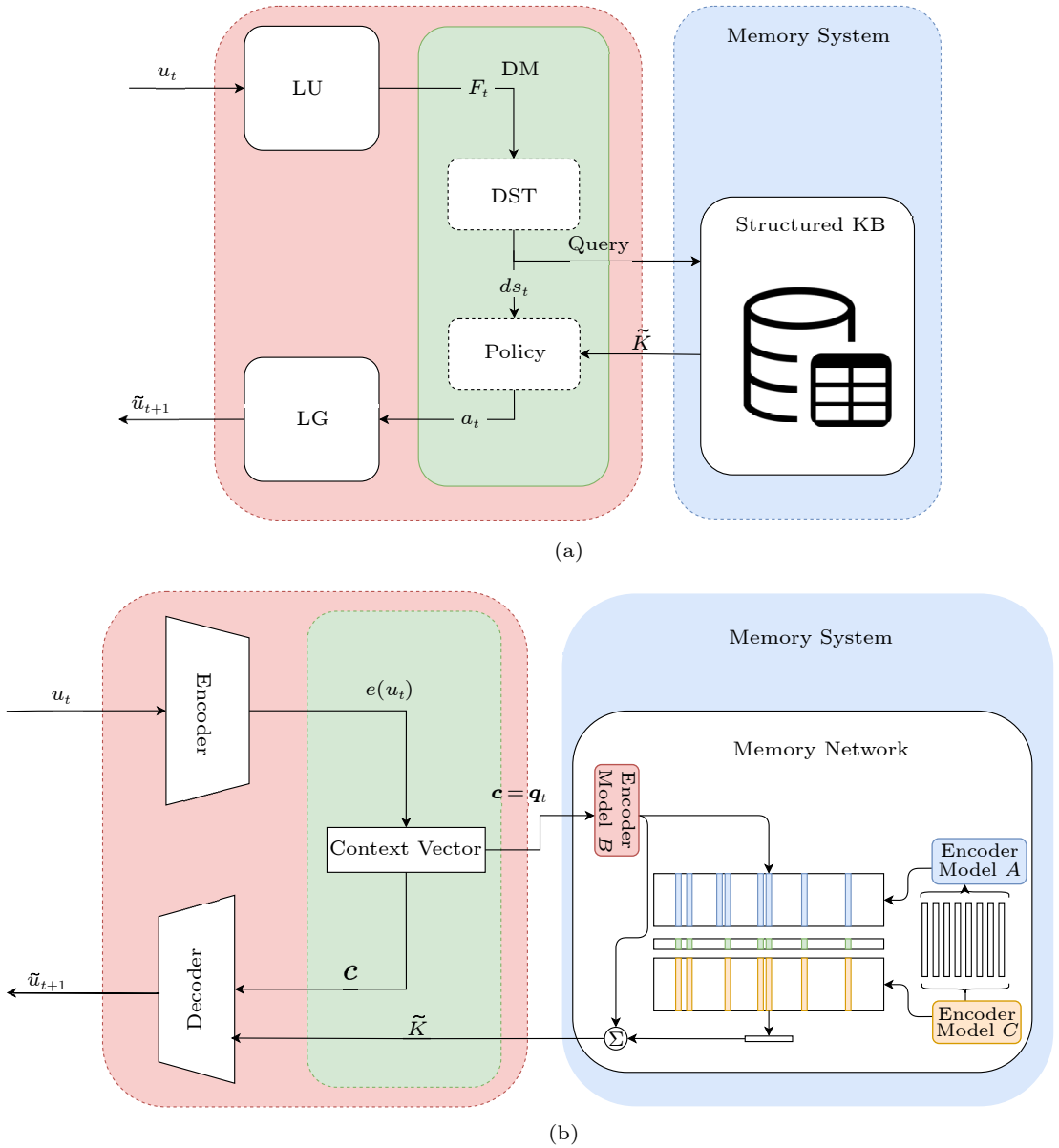

Fig.6. Two examples of implicit knowledge-enhancing interfaces. (a) Interface for the pipeline architecture. (b) Interface for the end-to-end architecture.

lows hierarchical reasoning on the input, stratifying the complexity of the KCA allowing, for example, to develop agents with different specialized dialogue systems that can be called at need. Generalizing this concept, the detection system could work as a driver that can sort the input requests to the most suitable sub-systems according to their requirements.

2) *Selection*

The selection task addresses the critical challenge of efficiently retrieving relevant knowledge snippets from a vast memory system for each knowledge-seeking turn. This task aims to provide the KCA with a curated subset of knowledge snippets, denoted as $\tilde{K} \subset \mathcal{M}$, which are contextually relevant to the ongoing dialogue context $U_{t,\,w}$. The goal is to enable KCAs to access and utilize pertinent information effectively enhancing their ability to provide informed and contextually accurate responses during conversations.

An accurate knowledge selection is crucial due to its impact on the perceived quality of the final response[40].

A common approach consists in considering couples (U_t, k_i), composed of the dialogue context U_t and a knowledge snippet k_i, and addressing the problem as a binary classification:

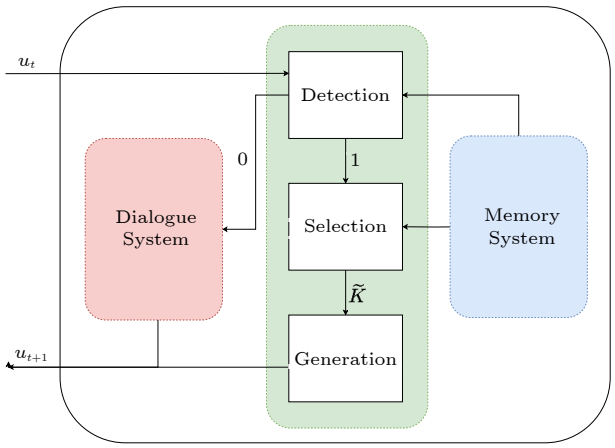

Fig.7. Architecture of an explicit knowledge-enhancing interface based on three subsystems.

$$f_S(U_t, k_i) = \begin{cases} 1, & \text{if } k_i \text{ is relevant for } U_t, \\ 0, & \text{else,} \end{cases}$$

in order to obtain a set of relevant knowledge snippets $\tilde{K}_t = \{k_i : f_S(U_t, k_i) = 1,\ \forall k_i \in K\}$.

The most common approach consists in computing a semantic similarity score[82] by mapping both U_t and k_i into fixed-length feature vectors using an encoder model e. Then, a scoring function S (e.g., the cosine similarity) is used to compute the similarity score $sim_{t,i}$ that represents how relevant the knowledge snippet k_i is to the current dialogue context U_t. The final classification can be performed by setting a threshold th and filtering only k_i for which $sim_{t,i}$ is greater than th or selecting only the top-k knowledge snippets based on their score.

Modern approaches are based on the so-called Dense Passage Retrieval[83] strategy that adopts pre-trained transformer models used as encoders. [84] proposes the Dense Knowledge Retrieval method, to efficiently select the relevant knowledge snippets in a dense embedding space.

An alternative approach consists in framing the task as a Question-Answering problem (Fig.8), where U_t is the question and k_i is a possible answer. This strategy is usually referred to as Passage Re-Ranking[85]. [40] proposes a BERT model that takes in input the concatenation of U_t and k_i separated by a [SEP] token. The output corresponding to the [CLS] token is then passed to a single-layer neural network to obtain the similarity score $sim_{t,i}$. Typically, these model are fine-tuned with a cross-entropy loss:

$$L = -\sum_{j \in J_{\text{pos}}} \log(s_j) - \sum_{j \in J_{\text{neg}}} \log(1 - s_j),$$

where J_{pos} is a subset of relevant knowledge snippets and J_{neg} is a subset of non-relevant knowledge snippets. [78] shows how the selection of negative examples J_{neg} during the training phase has a major impact on the ability of the model to filter relevant knowledge snippets. In particular, they design a procedure to select J_{neg} based on four steps with increasing difficulty. The snippets are selected: Random, In-Domain, In-Entity, and Cross-Entity (from entities aforementioned in the dialogue context). A major problem with this approach is that at inference time the model needs to check all the knowledge snippets in the knowledge base K since there is no information about which are the relevant knowledge snippets. Different filtering strategies can be adopted. [80] proposes a method named Retrieve&Rank that consists of a two-step approach: first, entity tracking is performed to retrieve all the entities named in the dialogue context that are then used to filter and select only related knowledge snippets; second, the knowledge ranking task described above is performed.

In the context of RAG, several optimizations were proposed to improve the retrieval process. [86] introduces BEQUE, a framework that utilizes LLMs for query rewriting, aiming to bridge the semantic gap, particularly for long-tail queries, and improve retrieval results. [87] addresses the challenge of creating effective zero-shot dense retrieval systems. It introduces hypothetical document embeddings (HyDE), which uses an instruction-following language model to generate a hypothetical document capturing relevance patterns, and then encodes it with an unsupervised contrastive encoder. The method involves manipulating queries to generate hypothetical documents that are then encoded into embedding vectors to identify similar real documents in the corpus.

3) *Generation*

Given a dialogue context $U_{t,w}$ and a set of relevant knowledge snippets $\tilde{K}_t$ with respect to u_t, the generation system aims to generate an appropriate system response u_{t+1} conditioned on the current dialogue context and the selected knowledge.

$$f_G(U_t, \tilde{K}_t) = u_{t+1}.$$

A common solution involves the fine-tuning of a transformer-based decoder model and incorporating the knowledge as supplementary context during the generation process. In this setting, during the training phase the dialogue context $U_{t,\,w}$ is passed along the knowledge snippets from $\tilde{K}_t$ and the model is let generate the output sequence u_{t+1}. To reduce the discrepancy between training and inference, word-level

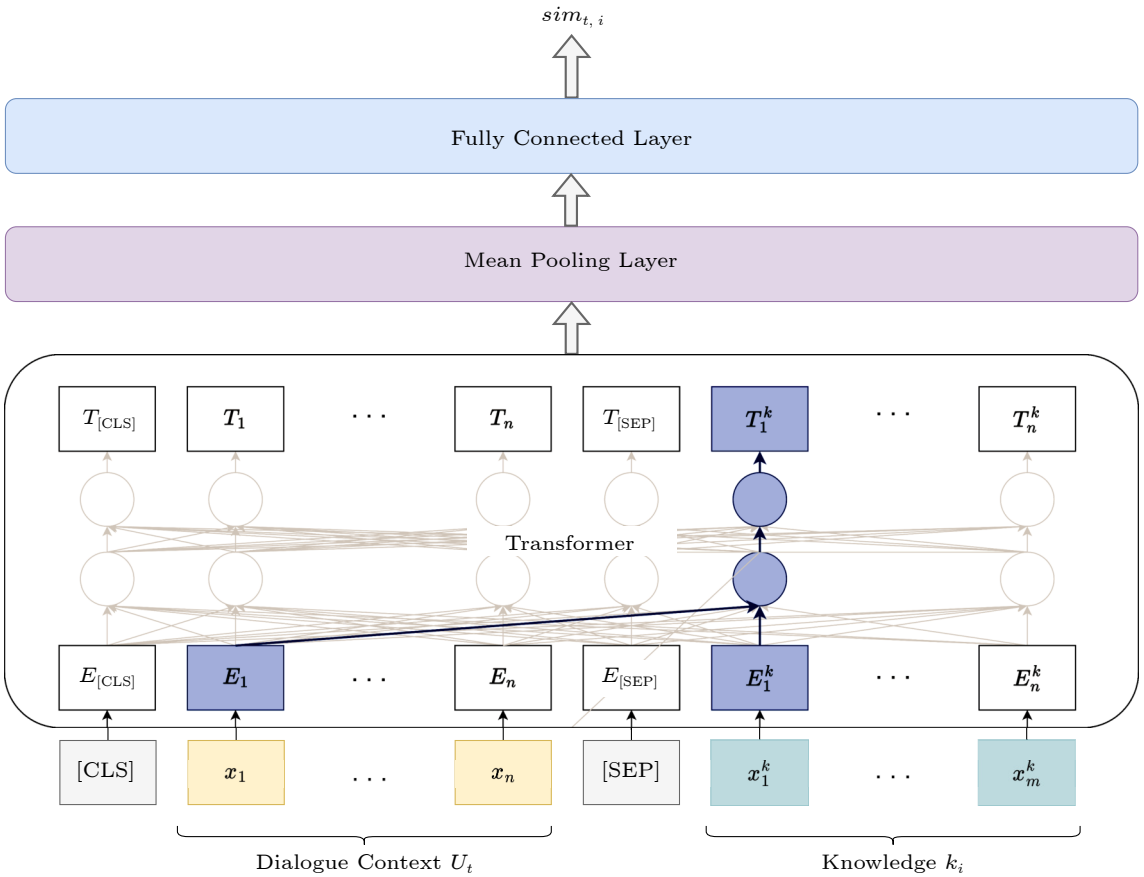

Fig.8. General structure for knowledge selection as a Q&A task using a Transformer-based model.

re-sampling techniques can be adopted[78, 88].

Alternatively to the simple generation of the tokens, copy mechanisms allow to choose sub-sequences in the input sequence and put them at proper places in the output sequence[89]. Considering $\mathcal{V}$ the global vocabulary and $\mathcal{X}$ the vocabulary of the distinct tokens from the input sequence, the idea consists in computing two scoring functions at each time step t:

- $\psi_g(y_t = v_i)$, $\forall v_i \in \mathcal{V}$, the score for generating token v_i,
- $\psi_c(y_t = x_i)$, $\forall x_i \in \mathcal{X}$, the score for copying token x_i,

or analogously computing the switching probability between the *copy* and *generate* mode[90]. This method can be exploited by extending the vocabulary with an additional vocabulary from the knowledge source to generate knowledge-enhanced text[91]. For example, [80] uses a latent variable to encode dialogue history and some selected knowledge, and generates the response combined with copy mechanism.

4 Applications (RQ3)

This section describes KCAs' different applications in four topic areas, focusing on the active usage of additional knowledge to provide more detailed services and improve users' experience.

4.1 Introduction

The global market of CAs is estimated to grow from USD 13.2 billion in 2024 to USD 49.9 by 2030 at a compound annual growth rate (CAGR) of almost 25%[14]. In a famous keynote speech from 2016, the CEO of Microsoft Satya Nadella affirmed that "Bots are the new apps" forecasting the ways humans will interact with machines in the future. Indeed, the modes users interact with Web contents and applications are changing moving towards more conversational forms. In the same year, chat and messaging apps surpassed by over 20% social network apps for a number of active users, becoming the first form of in-

[14] https://www.marketsandmarkets.com/Market-Reports/conversational-ai-market-49043506.html, May 2024.

formation exchange over the Internet[15]. Nowadays, users do not use messaging apps just to chat with friends but also to connect with brands, browse merchandise, and watch content[92].

For these reasons, more and more companies are integrating conversational AI technologies due to the many benefits that they bring with respect to traditional solutions.

4.1.1 Benefits of CAs

Minimizing the waiting time has been demonstrated to be a crucial aspect to customer satisfaction[93]. CAs provide full-time service availability for users, ensuring that assistance and information access are accessible at any time. This 24/7 availability reduces service downtime, offering users immediate support even outside regular business hours and eliminating long frustrating waiting times. CAs allow the possibility to deploy services across multiple channels offering a flexible and adaptable interface that accommodates various communication modalities. Users can interact with CAs through text, voice, or even multimodal interfaces, choosing the mode that suits their preferences and circumstances. This flexibility enhances user convenience and accessibility.

The scalability provided by conversational AI solutions can reduce operation costs automating workflows and freeing employees from repetitive tasks.

4.1.2 Benefits of KCAs

One of the most compelling advantages of KCAs empowered by additional knowledge resources is their capacity for personalization. These agents can craft highly individualized interactions, catering to the unique preferences, needs, and characteristics of each user. This major exploitation of contextual knowledge can help in designing more dynamic interfaces able to adaptively recommend the resources, offering personalized experiences based on the user's needs, and improving user engagement[94].

The additional knowledge can also contribute to developing more robust models able to deal with out-of-scope requests[40], providing frictionless interactions instead of relying on generic fallback intents such as "Sorry, I'm not able to help you with this." that will lead to a bad user experience. At the same time, the conversational interactions can be used to gather detailed feedback in order to improve furthermore the service provided.

4.2 General Assistance and Customer Service

General assistance and customer service are two of the most widely adopted fields for conversational agents. These were also the scopes for which the first task-oriented models were originally designed.

Early conversational agents used approaches based on static scripts defining apriori different possible flows and choosing at each step a predetermined path based on the input user query. Even if it is a bit outdated, this pattern still represents one of the most adopted strategies. Tools such as Google's Dialogflow[16] allow to design conversational agents without specialized programming skills.

The exploitation of knowledge can help design more sophisticated agents to dynamically retrieve product or resource information in databases and provide personalized assistance based on users' needs.

SamBot[95] is a KCA introducing an interactive consumer-centric learning approach to deliver appropriate answers to users' questions based on knowledge of Samsung marketing domain such as Samsung promotions, frequently asked questions (FAQs), and general knowledge. In the context of e-commerce, [96] presents SuperAgent, a KCA able to provide customer service leveraging additional knowledge such as products specifications, FAQ documents as well as user-generated content (e.g., product reviews). Similarly, RAGE[97] is an agent able to answer product-related questions based on knowledge collected directly from the users reviews of the corresponding product.

Considering assistance for financial services, [31] proposes StockBabble to support retail investors through a user-friendly conversational interface and supplementing their informational needs with additional knowledge about companies, trading recommendations, and the latest news. [98] introduces a novel method that combines RAG with a knowledge graph to improve customer service question answering. The method retains intra-issue structure and inter-issue relations, leading to better retrieval accuracy and answer quality. The method has been implemented within LinkedIn's customer service team, re-

[15] https://www.businessinsider.com/the-messaging-app-report-2016-4-23?r=US&IR=T, May 2024.
[16] https://cloud.google.com/dialogflow, May 2024.

ducing the median per-issue resolution time by 28.6%.

4.3 Healthcare

In healthcare, KCAs access extensive medical databases and the latest research findings to assist healthcare professionals in diagnosing conditions, recommending suitable treatment options, and educating patients about their health. Improving "Health Communication and Health Information Technology" was one of the Leading Health Indicators of the Healthy People 2020 (HP2020) program monitored by the National Center for Health Statistics[17]. [99] reports how the goal set by HP2020 to reach the 45% of the population able to access health information online without frustration[18] was not met. To this extent, KCAs can have a huge impact on healthcare providing an easy way to access and navigate through the myriad of information available on the Web or helping to perform a first self-assessment in non-critical scenarios, reducing the costs and improving the accessibility to care. An example in this sense is Babylon AI, a diagnostic system that allows people to automatically triage based on their self-reported symptoms, personal knowledge regarding their medical history and factual medical knowledge reviewed by clinical experts. This solution shows levels of safety and accuracy comparable to human doctors, demonstrating to be a valuable solution to save health workforce and resources[100].

[101] proposes iHelpr a simple chatbot that provides a guided self-assessment in the area of mental health. Afterward, based on this assessment it retrieves the most appropriate recommendations from its evidence-based knowledge base. In the case of escalated risk, it provides helpline numbers and emergency contact information to the user.

Among the over 70 000 Alexa skills available for Amazon Alexa, many are centered on improving mental and physical health. WebMD[19] for example, provides assistance giving answers to health-related questions such as symptoms, treatments, causes for conditions and definitions of medical terms. Similarly, [102] discusses the development of Almanac, a large language model framework designed for clinical medicine. The LLM model is augmented with retrieval capabilities to provide medical guideline and treatment recommendations. Almanac was evaluated on a novel dataset of clinical scenarios by a panel of physicians showing significant improvements in factuality, completeness, and safety across various specialties.

All these models operate reactively, taking action only after being interrogated. A goal for future KCAs is to operate proactively, providing dynamic assistance through user-specific suggestions or reminders. To this extent, KCAs equipped with knowledge from medical databases can provide patients with personalized health advice, taking into account their medical history, current symptoms, and relevant clinical guidelines. Preventive intervention for modifiable risk factors is an interesting area of research for KCAs in healthcare. [103] introduces Health-LLM, a personalized disease prediction model that integrates large-scale feature extraction with medical knowledge. Health-LLM offers detailed task information from health reports, professional medical expertise adjustment, and semi-automated feature extraction to enhance disease prediction accuracy. Health-LLM outperforms traditional methods and other large language models in accuracy and F1 score, demonstrating its potential to improve disease prediction and personalized health management.

One of the primary limitations preventing the widespread adoption of LLMs in the healthcare sector is the concern that even robust LLMs may generate responses that are unsuitable for use in the safety-critical medical domain[11]. The enhancement of these models with external knowledge presents a promising solution. [104] proposes an innovative approach that leverages a medical knowledge graph derived from the National Library of Medicine's Unified Medical Language System (UMLS) to augment the proficiency of LLMs in automated diagnosis generation. The incorporation of this KG allows for the interpretation and summarization of complex medical concepts. [105] introduces SourceCheckup, an evaluation framework to score LLMs' ability to cite relevant medical sources. The framework was validated against a panel of medical doctors, showing 88% agreement with GPT-4's source relevance validation. However, it was found

[17] https://www.cdc.gov/nchs/healthy_people/hp2020.htm, May 2024.

[18] Accessing health information without frustration is defined as responding *"strongly agree"* to the question: *"Based on the results of your most recent search for information about health or medical topics, how much do you agree or disagree with the following statement? You felt frustrated during your search for the information."* Response options were measured on a 4-point scale, ranging from 1 = strongly disagree to 4 = strongly agree[99].

[19] https://www.amazon.com/WebMD-Health-Corp/dp/B01MRM361G, May 2024.

that 50% to 90% of the statements of GPT-4, and about 30% of the statements of GPT-4 plus RAG knowledge augmentation, are not fully supported. This study highlights the necessity of further improvements to fortify the reliability of these models in critical environments.

4.4 Education

In traditional learning environments, it is difficult to provide personalized learning support and to improve the learning experience based on the specific learner's personal needs. Growing research interest has been displayed in pedagogical conversational agents[106, 107]. In education, KCAs integrate external educational resources, textbooks, and reference materials, can assist students with homework, answer academic queries, and explain complex concepts, ultimately fostering a conducive learning environment. Moreover, they can tailor their guidance to a student's learning pace and individual challenges, ensuring a more effective and personalized learning journey[108][20].

[35] proposes the Personal Assistant for Life-Long Learning PAL3 with the goal to create an agent that can accompany learners throughout their academic careers. It is implemented as a virtual embodied KCA with an adaptive recommendation engine that provides personalized access to a set of knowledge resources and learning material. This agent prevents skill decay in between periods of formal instruction to monitor and increase the engagement and motivation of the student through the gamification of the learning process. At the moment the prototype is restricted only to a single domain, supporting US Navy sailors to advance from their level of electronics technician training. However, ensuring accurate personalization requires reliable knowledge sources and an understanding of diverse learning profiles.

Considering the universally recognized key role that conversational interaction has in learning a second language[109], the adoption of KCAs to this extent seems to be fairly natural. KCAs can facilitate language learning by providing real-time translations, language exercises, pronunciation guidance, and cultural context from external knowledge.

In 2023, Duolingo introduced Duolingo Max[21], an advanced subscription tier that leverages GPT-4 technology to enrich language learning with features like "Explain My Answer" and "Roleplay"[22], providing personalized feedback and interactive conversation practice to enhance the educational experience.

Finally, [110] adopts an interesting approach. In contrast with the majority of the studies, which use agents as peers or that make agents to imitate teachers, the authors overturned the roles of the two investigating the adoption of a KCA to support the learning-by-teaching paradigm, where the agent receives instructions from the students. The model called Curiosity Notebook allows students to select a topic and the model provides related articles sampled from a knowledge source in its memory system. Students can use a textual interface to highlight important sentences. After they finish, they can start a conversational interaction with the agent in which the agent will ask questions related to the document and the student tries to answer them. This interaction allows the students to reflect on their own knowledge and, as a result, to obtain a deeper comprehension of the sections on which they are most unprepared.

4.5 Games

One of the key aspects of a videogame is the immersion that it provides to the players. Nowadays, computer graphic techniques allow to recreate photorealistic virtual words with incredible richness of details. However, it is commonly agreed that the total photo and audio-realism are not necessary for a virtual environment to produce in the viewer a sense of immersion[111].

Another way to enhance the immersion is to increase the player's freedom to interact with the environment giving the feeling of a living dynamic world where the player's actions have repercussions. A classical example is providing a non-linear experience where the decision made by the player influences the progress of the story. In particular, one of the main channels through which the main story is carried forward is constituted by the interactions between the player and non-player characters (NPCs). However, little progress has been made in the last years to improve the techniques adopted to implement dynamic conversations in videogames. In modern games that

[20] https://blog.duolingo.com/learning-how-to-help-you-learn-introducing-birdbrain, May 2024.
[21] https://blog.duolingo.com/duolingo-max, May 2024.
[22] https://blog.duolingo.com/chatbot-language-practice, May 2024.

allow multiple choices, typically the dialogues with NPCs are highly structured and implemented through rule-based approaches for NLU that rely on a dialogue tree (or dialogue flow) where at each turn the player can select one option among a predefined set. The application of KCAs in this context can greatly increase the engagement of the players, letting the user have active participation in the interaction and providing conversational feedbacks to the players, tailored to their personal actions, making every player's experience with the game unique.

Façade[112] is an early experiment on this field. It is a real-time, first-person game that adopts a KCA to bring on an interactive drama. Recently, NVIDIA introduced Avatar Cloud Engine (ACE), a suite of technologies for creating lifelike digital avatars using generative AI (Fig.9). ACE transforms generic NPCs into interactive characters that can converse and assist players in games[23]. Similarly, Ubisoft presents NEO NPC, a generative AI prototype to enhance NPC interactions in video games. The goal is to create NPCs that players can have real conversations with, beyond pre-determined dialogue trees, maintaining the authenticity of the character and scenario. Writers craft the NPCs' backstories and personalities, which are then used as additional knowledge for the LLM. The model uses an iterative approach to refining the AI's dialogue, ensuring it stays true to the character's envisioned personality and reacts appropriately to player inputs[24].

Another interesting application for KCAs in the field of games is the text adventures in which players use textual commands to control the characters and influence the environment. AI Dungeon[25] is a text adventure game based on GPT-3 that exploits additional knowledge to create entertaining stories, retain information about the users profiles and their decisions, and model the background information about the environment and the events that already happened in the story in order to provide a coherent narration. The agent allows also the possibility to the user to dynamically intervene to modify the game environment by adding new knowledge.

5 Future Challenges (RQ4)

The refinement of knowledge-enhancing techniques for conversational agents is shown to be a promising research path to design agents able to provide more specific responses. To this extent many

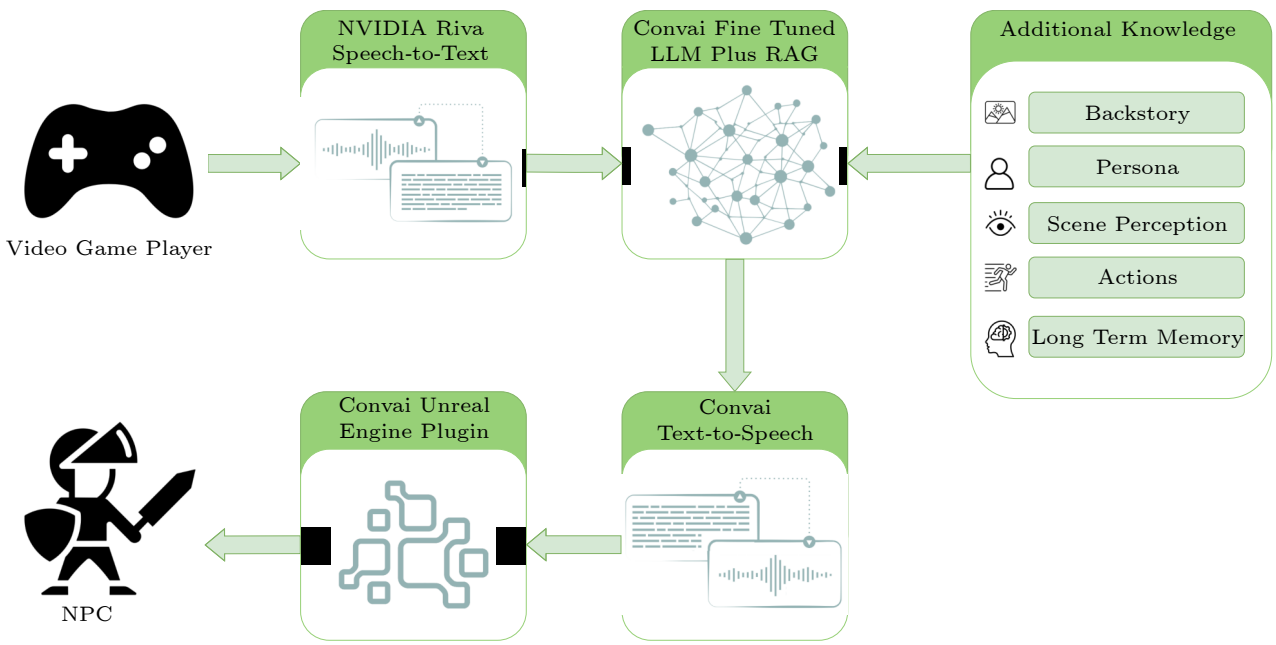

Fig.9. Example of the architecture proposed by NVIDIA ACE. The LLM is enhanced with additional contextual knowledge such as the character "Persona" and the background story in order to provide contextualized answers.

[23] https://www.nvidia.com/en-us/geforce/news/nvidia-ace-architecture-ai-npc-personalities, May 2024.
[24] https://news.ubisoft.com/en-us/article/5qXdxhshJBXoanFZApdG3L/how-ubisofts-new-generative-ai-prototype-changes-the-narrative-for-npcs, May 2024.
[25] https://play.aidungeon.com, May 2024.

progresses are still required. This section discusses different interesting paths of research for future developments of knowledge-enhanced conversational agents.

5.1 Few-Shot and Zero-Shot Learning

Knowledge-enhancing techniques can improve the performance of CAs in knowledge-intensive tasks. However, these solutions are typically task-specific and rely heavily on labeled data for training. Therefore, there is a pressing need for KCAs to enhance their ability to generalize their knowledge and skills, allowing them to swiftly adapt to new domains, in few or even zero-shot learning settings.

The schema-guided paradigm provides an interesting possible solution to dynamically scale to new domain making predictions over a dynamic set of intents and slots, provided as input, using their natural language descriptions[113]. Another promising path of research regards the application of meta-learning[114].

Finally, while LLMs have shown promise as few-shot learners in open-domain settings[4], their application in high-risk domains remains an open problem due to the challenge of reliably generating outputs[11].

5.2 Lifelong Learning

To be really efficient, KCAs should also be capable of learning continuously, gaining experience, and accumulating in their memory systems the knowledge acquired during the various interactions. This problem is referred to as continuous or lifelong learning[115]. Research in this area is still in the very early stages. [116] proposes a model for KCAs to enable them to interactively learn new knowledge during the conversations. Another interesting area of research focuses on the capabilities of the models to adapt to changing environments. This problem is significantly more difficult when new domains come with limited or zero training samples. [117] investigates adaptation to dynamic knowledge graphs. These graphs feature temporal states/entities and evolving relations, exploited to help KCAs to adapt seamlessly to shifting knowledge landscapes. Additionally, [118] explores the possibility of KCAs acquiring knowledge through self-learning in dialogues. This interdisciplinary work aims at developing self-learning conversational agents (SLCAs) capable of autonomously acquiring knowledge through interactions with individuals. The findings of this research promise significant theoretical and practical implications across various scientific domains.

However, lifelong learning presents several substantial challenges. One of the most important is the potential for knowledge forgetting over time, especially when faced with a continuous stream of new information. This phenomenon, often referred to as "catastrophic forgetting", requires innovative solutions to allow KCAs to retain and build upon previously acquired knowledge[119].

In summary, the pursuit of lifelong learning and adaptive capabilities, along with addressing challenges like catastrophic forgetting, will empower KCAs to remain relevant, knowledgeable, and effective in dynamic environments. These advancements will ensure that KCAs can consistently provide valuable support.

5.3 Reasoning System

The human language system has evolved to support efficient communications not to construct complex thoughts[120]. In particular, the human language system does not support non-linguistic cognitive abilities, but it is intimately linked with the system that supports social cognition. These discoveries suggest a dichotomous system composed of a language system, which handles the communication part and a reasoning system that deals with the sociability, integrating contextual information from the environment.

Concerning language systems, current neural language models are already quite advanced. Indeed, the human language system is fundamentally predictive, and modern neural language models based on next-word prediction (e.g., GPT-2) are able to predict nearly 100% of explainable variance in neural responses to sentences[121].

In this sense, further improvements in the conversational capabilities of KCAs would come from not only the advancement in language generation models, but also the development of proper reasoning systems able to cope with the social cognition part of the agents. In this setting, a reasoning system should be able to perform high-level inferences on the conversational input exploiting the knowledge already acquired and stored in the memory system and integrating it with contextual information coming from the environment such as situational factors and social cues[122].

In addition, a reasoning system could also implement subsystems to keep long-term goals in the conversation[123], analyze the input through conceptual models to identify structural wrongness in the input arguments[124].

5.4 Trustworthiness

One of the main challenges limiting the diffusion of conversational agents, especially in high-risk domains, is related to security and privacy issues that directly affect users' confidence and consequently their willingness to adopt the technology[125, 126].

Recent surveys underscore this growing awareness among users about their personal data. A survey of 1000 U.S. consumers revealed that 85% of respondents would not forgive a company's misuse of their data, even if they had previously placed trust in the brand[26]. Moreover, a staggering 91% expressed the desire for stringent government regulations to protect their data. The lack of trustworthiness can lead to reluctance among many individuals to engage in conversational tasks that involve sharing sensitive personal data[20].

This behavior shows an increasing demand for trustworthiness that can be achieved with improvements in the perceived quality in terms of usability, perceived security, and privacy of the data shared with the agents[127]. In this sense, further advancements are still necessary, especially in sectors that deal with confidential data such as banking[128], healthcare[129], and data about underage users[130].

In conclusion, the assurance of trustworthiness remains a pivotal objective in the evolution of conversational agents. Addressing security and privacy concerns, and increasing user perceived quality are central to the acceptance of these agents across various sectors and applications[131].

6 Conclusions

In this work, we outlined the technology supporting knowledge-enhanced conversational agents (KCAs), providing an extensive overview of the implementation techniques and reviewing different applications and possible paths for future developments.

The potential disruptiveness of KCAs is drawing interest from industries and the general public, and the related market is expected to explode in the following years with a compound annual growth rate (CAGR) of over 25%.

From a general point of view, KCAs position themselves as access points to resources or services, offering a valid alternative to more traditional solutions and providing a natural interface based on conversational features for the users.

The focal point of KCAs is the exploitation of different knowledge resources, such as domain-specific knowledge (e.g., product specifics), information about the user's profile, and contextual data (e.g., time and location) as a means to design highly personalized experiences tailored to each user's exigencies and thus leading to higher user engagement rates.

Despite the enormous potential of KCAs, many improvements are still necessary in order to enable the diffusion of these models. In particular, one of the main issues that are currently limiting their adoption is the perceived quality of the service. In this sense, a major challenge for the future regards a fair and transparent use of the data in order to gain more trustworthiness.

Conflict of Interest The authors declare that they have no conflict of interest.

References

[1] Mayor A. Gods and Robots. Princeton University Press, 2018.

[2] Weizenbaum J. ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 1966, 9(1): 36–45. DOI: 10.1145/365153.365168.

[3] Hofstadter D R. Preface 4 The ineradicable Eliza effect and its dangers, epilogue. In *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*, Hofstadter D R (ed.), Basic Books, Inc., 1996, p.157.

[4] Brown T B, Mann B, Ryder N *et al.* Language models are few-shot learners. In *Proc. the 34th Int. Conf. Neural Information Processing Systems*, Dec. 2020, Article No. 159.

[5] Ouyang L, Wu J, Jiang X *et al.* Training language models to follow instructions with human feedback. In *Proc. the 36th Int. Conf. Neural Information Processing Systems*, Nov. 2022, Article No. 2011.

[6] Kovács B. The Turing test of online reviews: Can we tell the difference between human-written and GPT-4-written online reviews? *Marketing Letters*, 2024. DOI: 10.

[26] https://www.forbes.com/sites/gilpress/2019/11/25/ai-stats-news-chatbots-increase-sales-by-67-but-87-of-consumers-prefer-humans, May 2024.

1007/s11002-024-09729-3.

[7] Bender E M, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? In *Proc. the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Mar. 2021, pp.610–623. DOI: 10.1145/3442188.3445922.

[8] Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, Steinhardt J. Measuring massive multitask language understanding. arXiv: 2009.03300, 2021. https://arxiv.org/abs/2009.03300, May 2024.

[9] Zhang Y, Li Y F, Cui L Y *et al.* Siren's song in the AI ocean: A survey on hallucination in large language models. arXiv: 2309.01219, 2023. https://arxiv.org/abs/2309.01219, May 2024.

[10] Kandpal N, Deng H K, Roberts A, Wallace E, Raffel C. Large language models struggle to learn long-tail knowledge. In *Proc. the 40th International Conference on Machine Learning*, Jul. 2023, pp.15696–15707.

[11] Singhal K, Azizi S, Tu T *et al.* Large language models encode clinical knowledge. *Nature*, 2023, 620(7972): 172–180. DOI: 10.1038/s41586-023-06291-2.

[12] Eggins S, Slade D. Analysing Casual Conversation. Equinox, 2004.

[13] Lewis P, Perez E, Piktus A *et al.* Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proc. the 34th International Conference on Neural Information Processing Systems*, Dec. 2020, Article No.793.

[14] Yang L F, Chen H Y, Li Z, Ding X, Wu X D. Give us the facts: Enhancing large language models with knowledge graphs for fact-aware language modeling. *IEEE Trans. Knowledge and Data Engineering*, 2024, 36(7): 3091–3110. DOI: 10.1109/TKDE.2024.3360454.

[15] Syvänen S, Valentini C. Conversational agents in online organization–stakeholder interactions: A state-of-the-art analysis and implications for further research. *Journal of Communication Management*, 2020, 24(4): 339–362. DOI: 10.1108/JCOM-11-2019-0145.

[16] Ram A, Prasad R, Khatri C *et al.* Conversational AI: The science behind the Alexa prize. arXiv: 1801.03604, 2018. https://arxiv.org/abs/1801.03604, May 2024.

[17] Chen H S, Liu X R, Yin D W, Tang J L. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter*, 2017, 19(2): 25–35. DOI: 10.1145/3166054.3166058.

[18] Huang M L, Zhu X Y, Gao J F. Challenges in building intelligent open-domain dialog systems. *ACM Trans. Information Systems*, 2020, 38(3): Article No. 21. DOI: 10.1145/3383123.

[19] Hussain S, Ameri Sianaki O, Ababneh N. A survey on conversational agents/chatbots classification and design techniques. In *Proc. the 33rd Int. Conf. Advanced Information Networking and Applications*, Mar. 2019, pp.946–956. DOI: 10.1007/978-3-030-15035-8_93.

[20] de Barcelos Silva A, Gomes M M, da Costa C A, da Rosa Righi R, Barbosa J L V, Pessin G, De Doncker G, Federizzi G. Intelligent personal assistants: A systematic literature review. *Expert Systems with Applications*, 2020, 147: 113193. DOI: 10.1016/j.eswa.2020.113193.

[21] Gemini Team. Gemini: A family of highly capable multimodal models. arXiv: 2312.11805, 2023. https://arxiv.org/abs/2312.11805, May 2024.

[22] Bao S Q, He H, Wang F, Wu H, Wang H F, Wu W Q, Guo Z, Liu Z B, Xu X C. PLATO-2: Towards building an open-domain chatbot via curriculum learning. In *Proc. the 2021 Findings of the Association for Computational Linguistics*, Aug. 2021, pp.2513–2525. DOI: 10.18653/v1/2021.findings-acl.222.

[23] Adiwardana D, Luong M T, So D R, Hall J, Fiedel N, Thoppilan R, Yang Z, Kulshreshtha A, Nemade G, Lu Y F, Le Q V. Towards a human-like open-domain chatbot. arXiv: 2001.09977, 2020. https://arxiv.org/abs/2001.09977, May 2024.

[24] Papaioannou I, Cercas Curry A, Part J *et al.* Alana: Social dialogue using an ensemble model and a ranker trained on user feedback. In *Proc. the 2017 Alexa Prize*, Aug. 2017.

[25] Ferrara E, Varol O, Davis C, Menczer F, Flammini A. The rise of social bots. *Communications of the ACM*, 2016, 59(7): 96–104. DOI: 10.1145/2818717.

[26] Subrahmanian V S, Azaria A, Durst S, Kagan V, Galstyan A, Lerman K, Zhu L H, Ferrara E, Flammini A, Menczer F. The DARPA twitter bot challenge. *Computer*, 2016, 49(6): 38–46. DOI: 10.1109/MC.2016.183.

[27] Roller S, Dinan E, Goyal N *et al.* Recipes for building an open-domain chatbot. In *Proc. the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Apr. 2021, pp.300–325. DOI: 10.18653/v1/2021.eacl-main.24.

[28] Sun H T, Dhingra B, Zaheer M *et al.* Open domain question answering using early fusion of knowledge bases and text. In *Proc. the 2018 Conference on Empirical Methods in Natural Language Processing*, Oct. 31-Nov. 1, 2018, pp.4231–4242. DOI: 10.18653/v1/D18-1455.

[29] Porcheron M, Fischer J E, Reeves S, Sharples S. Voice interfaces in everyday life. In *Proc. the 2018 CHI Conference on Human Factors in Computing Systems*, Apr. 2018, Article No. 640. DOI: 10.1145/3173574.3174214.

[30] Pieraccini R. The Voice in the Machine: Building Computers that Understand Speech. MIT Press, 2012.

[31] Sharma S, Brennan J, Nurse J. StockBabble: A conversational financial agent to support stock market investors. In *Proc. the 3rd Conference on Conversational User Interfaces*, Jul. 2021, Article No. 25. DOI: 10.1145/3469595.3469620.

[32] Fraser J, Papaioannou I, Lemon O. Spoken conversational AI in video games: Emotional dialogue management increases user engagement. In *Proc. the 18th International Conference on Intelligent Virtual Agents*, Nov. 2018, pp.179–184. DOI: 10.1145/3267851.3267896.

[33] McTear M, Callejas Z, Griol D. The Conversational Interface: Talking to Smart Devices. Springer, 2016. DOI: 10.1007/978-3-319-32967-3.

[34] Sarikaya R. The technology behind personal digital assistants: An overview of the system architecture and key components. *IEEE Signal Processing Magazine*, 2017, 34(1): 67–81. DOI: 10.1109/MSP.2016.2617341.

[35] Swartout W R, Nye B D, Hartholt A *et al*. Designing a personal assistant for life-long learning (PAL3). In *Proc. the 29th International Florida Artificial Intelligence Research Society Conference*, May 2016, pp.491–496.

[36] Cicció J A, Quesada L. Framework for creating audio games for intelligent personal assistants. In *Proc. the 2017 International Conference on Advances in Human Factors and Wearable Technologies*, July 2017, pp.204–214. DOI: 10.1007/978-3-319-60639-2_21.

[37] McTear M. Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots. Springer, 2020: 1–251. DOI: 10.1007/978-3-031-02176-3.

[38] Minsky M. Society of mind: A response to four reviews. *Artificial Intelligence*, 1991, 48(3): 371–396. DOI: 10.1016/0004-3702(91)90036-J.

[39] Marková I, Linell P, Grossen M, Salazar Orvig A. Dialogue in Focus Groups: Exploring Socially Shared Knowledge. Equinox Publishing, 2007.

[40] Kim S, Eric M, Hedayatnia B, Gopalakrishnan K, Liu Y, Huang C W, Hakkani-Tur D. Beyond domain APIs: Task-oriented conversational modeling with unstructured knowledge access track in DSTC9. arXiv: 2101.09276, 2021. https://arxiv.org/abs/2101.09276, May 2024.

[41] Chen Q, Zhuo Z, Wang W. BERT for joint intent classification and slot filling. arXiv: 1902.10909, 2019. https://arxiv.org/abs/1902.10909, May 2024.

[42] Castellucci G, Bellomaria V, Favalli A, Romagnoli R. Multi-lingual intent detection and slot filling in a joint BERT-based model. arXiv: 1907.02884, 2019. https://arxiv.org/abs/1907.02884, May 2024.

[43] Wang J X, Wei K, Radfar M, Zhang W W, Chung C. Encoding syntactic knowledge in transformer encoder for intent detection and slot filling. In *Proc. the 35th AAAI Conference on Artificial Intelligence*, Feb. 2021, pp.13943–13951. DOI: 10.1609/aaai.v35i16.17642.

[44] Zhang J G, Hashimoto K, Wu C S, Wang Y, Yu P, Socher R, Xiong C M. Find or classify? Dual strategy for slot-value predictions on multi-domain dialog state tracking. In *Proc. the 9th Joint Conference on Lexical and Computational Semantics*, Dec. 2020, pp.154–167.

[45] Gao S Y, Sethi A, Agarwal S, Chung T, Hakkani-Tür D Z. Dialog state tracking: A neural reading comprehension approach. In *Proc. the 20th Annual SIGdial Meeting on Discourse and Dialogue*, Sept. 2019, pp.264–273. DOI: 10.18653/v1/W19-5932.

[46] Zhou L, Small K. Multi-domain dialogue state tracking as dynamic knowledge graph enhanced question answering. arXiv: 1911.06192, 2019. https://arxiv.org/abs/1911.06192, May 2024.

[47] Rohmatillah M, Chien J T. Advances and challenges in multi-domain task-oriented dialogue policy optimization. *APSIPA Trans. Signal and Information Processing*, 2023, 12(1): e37. DOI: 10.1561/116.00000132.

[48] Peng B L, Zhu C G, Li C Y, Li X J, Li J C, Zeng M, Gao J F. Few-shot natural language generation for task-oriented dialog. arXiv: 2002.12328, 2020. https://arxiv.org/abs/2002.12328, May 2024.

[49] Richards B A, Lillicrap T P. Dendritic solutions to the credit assignment problem. *Current Opinion in Neurobiology*, 2019, 54: 28–36. DOI: 10.1016/j.conb.2018.08.003.

[50] Gao J F, Galley M, Li L H. Neural approaches to conversational AI. In *Proc. the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, Jul. 2018, pp.1371–1374. DOI: 10.1145/3209978.3210183.

[51] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks. In *Proc. the 27th International Conference on Neural Information Processing Systems*, Dec. 2014, pp.3104–3112.

[52] Clark K, Luong M T, Le Q V, Manning C D. ELECTRA: Pre-training text encoders as discriminators rather than generators. arXiv: 2003.10555, 2020. https://arxiv.org/abs/2003.10555, May 2024.

[53] Song K T, Tan X, Qin T, Lu J F, Liu T Y. MPNet: Masked and permuted pre-training for language understanding. In *Proc. the 34th Int. Conf. Neural Information Processing Systems*, Dec. 2020, Article No.1414.

[54] Gao T Y, Yao X C, Chen D Q. SimCSE: Simple contrastive learning of sentence embeddings. In *Proc. the 2021 Conference on Empirical Methods in Natural Language Processing*, Nov. 2021, pp.6894–6910. DOI: 10.18653/v1/2021.emnlp-main.552.

[55] Chowdhery A, Narang S, Devlin J *et al*. PaLM: Scaling language modeling with pathways. *The Journal of Machine Learning Research*, 2022, 24(1): 240.

[56] Petroni F, Rocktäschel T, Riedel S, Lewis P, Bakhtin A, Wu Y X, Miller A. Language models as knowledge bases? In *Proc. the 2019 EMNLP-IJCNLP*, Nov. 2019, pp.2463–2473. DOI: 10.18653/v1/D19-1250.

[57] Taylor R, Kardas M, Cucurull G, Scialom T, Hartshorn A, Saravia E, Poulton A, Kerkez V, Stojnic R. Galactica: A large language model for science. arXiv: 2211.09085, 2022. https://arxiv.org/abs/2211.09085, May 2024.

[58] Yu W H, Zhu C G, Li Z T, Hu Z T, Wang Q Y, Ji H, Jiang M. A survey of knowledge-enhanced text generation. *ACM Computing Surveys*, 2022, 54(11s): 227. DOI: 10.1145/3512467.

[59] Adamopoulou E, Moussiades L. Chatbots: History, technology, and applications. *Machine Learning with Applications*, 2020, 2: 100006. DOI: 10.1016/j.mlwa.2020.100006.

[60] Huang T H, Chang J C, Bigham J P. Evorus: A crowd-powered conversational assistant built to automate itself over time. In *Proc. the 2018 CHI Conference on Human Factors in Computing Systems*, Apr. 2018, Article No.295. DOI: 10.1145/3173574.3173869.

[61] Liu D Y H, Yan Y, Gong Y Y *et al*. GLGE: A new general language generation evaluation benchmark. In *Proc. the 2021 Findings of the Association for Computational Linguistics*, Aug. 2021, pp.408–420. DOI: 10.18653/v1/2021.findings-acl.36.

[62] Ji S X, Pan S R, Cambria E, Marttinen P, Yu P S. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Trans. Neural Networks and Learning Systems*, 2022, 33(2): 494–514. DOI: 10.1109/TNNLS.2021.3070843.

[63] Young T, Cambria E, Chaturvedi I, Zhou H, Biswas S,

Huang M L. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Proc. the 32nd AAAI Conference on Artificial Intelligence*, Feb. 2018, pp.4970–4977. DOI: 10.1609/aaai.v32i1.11923.

[64] Zhang H Y, Liu Z H, Xiong C Y, Liu Z Y. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In *Proc. the 58th Annual Meeting of the Association for Computational Linguistics*, Jul. 2020, pp.2031–2043. DOI: 10.18653/v1/2020.acl-main.184.

[65] Yang S Q, Zhang R, Erfani S. GraphDialog: Integrating graph knowledge into end-to-end task-oriented dialogue systems. In *Proc. the 2020 EMNLP*, Nov. 2020, pp.1878–1888. DOI: 10.18653/v1/2020.emnlp-main.147.

[66] Ghazvininejad M, Brockett C, Chang M W, Dolan B, Gao J F, Yih W T, Galley M. A knowledge-grounded neural conversation model. In *Proc. the 32nd AAAI Conference on Artificial Intelligence*, Feb. 2018, pp.5110–5117. DOI: 10.1609/aaai.v32i1.11977.

[67] Weston J, Chopra S, Bordes A. Memory networks. arXiv: 1410.3916, 2015. https://arxiv.org/abs/1410.3916, May 2024.

[68] Sukhbaatar S, Szlam A, Weston J *et al.* End-to-end memory networks. In *Proc. the 28th Int. Conf. Neural Information Processing Systems*, Dec. 2015, pp.2440–2448.

[69] Miller A, Fisch A, Dodge J, Karimi A H, Bordes A, Weston J. Key-value memory networks for directly reading documents. In *Proc. the 2016 Conference on Empirical Methods in Natural Language Processing*, Nov. 2016, pp.1400–1409. DOI: 10.18653/v1/D16-1147.

[70] Vougiouklis P, Hare J, Simperl E. A neural network approach for knowledge-driven response generation. In *Proc. the 26th Int. Conf. Computational Linguistics*: *Technical Papers*, Dec. 2016, pp.3370–3380.

[71] Madotto A, Wu C S, Fung P. Mem2Seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *Proc. the 56th Annual Meeting of the Association for Computational Linguistics (Volume* 1), Jul. 2018, pp.1468–1478. DOI: 10.18653/v1/P18-1136.

[72] Gangi Reddy R, Contractor D, Raghu D, Joshi S. Multi-level memory for task oriented dialogs. In *Proc. the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, *(Volume* 1), Jun. 2019, pp.3744–3754. DOI: 10.18653/v1/N19-1375.

[73] Wen J T, Jiang D Z, Tu G, Liu C, Cambria E. Dynamic interactive multiview memory network for emotion recognition in conversation. *Information Fusion*, 2023, 91: 123–133. DOI: 10.1016/j.inffus.2022.10.009.

[74] Zhang Z, Takanobu R, Zhu Q, Huang M L, Zhu X Y. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, 2020, 63(10): 2011–2027. DOI: 10.1007/s11431-020-1692-3.

[75] Koncel-Kedziorski R, Bekal D, Luan Y *et al.* Text generation from knowledge graphs with graph transformers. In *Proc. the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, *(Volume* 1), Jun. 2019, pp.2284–2293. DOI: 10.18653/v1/N19-1238.

[76] Fu Y, Feng Y S. Natural answer generation with heterogeneous memory. In *Proc. the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, *(Volume* 1), Jun. 2018, pp.185–195. DOI: 10.18653/v1/N18-1017.

[77] Mi H T, Ren Q Y, Dai Y P, He Y F, Sun J, Li Y B, Zheng J, Xu P. Towards generalized models for beyond domain API task-oriented dialogue. In *Proc. the 2021 AAAI-21 DSTC9 Workshop*, Feb. 2021.

[78] He H, Lu H, Bao S Q, Wang F, Wu H, Niu Z Y, Wang H F. Learning to select external knowledge with multi-scale negative sampling. *IEEE/ACM Trans. Audio, Speech, and Language Processing*, 2024, 32: 714–720. DOI: 10.1109/TASLP.2023.3301222.

[79] Jin D, Gao S Y, Kim S, Liu Y, Hakkani-Tur D. Towards zero and few-shot knowledge-seeking turn detection in task-orientated dialogue systems. In *Proc. the 3rd Workshop on Natural Language Processing for Conversational AI*, Nov. 2021, pp.281–288. DOI: 10.18653/v1/2021.nlp4convai-1.27.

[80] Tan C H, Yang X Y, Zheng Z O, Li T D, Feng Y F, Gu J C, Liu Q, Liu D, Ling Z H, Zhu X D. Learning to retrieve entity-aware knowledge and generate responses with copy mechanism for task-oriented dialogue systems. arXiv: 2012.11937, 2020. https://arxiv.org/abs/2012.11937, May 2024.

[81] Wang Y L, Li P, Sun M S, Liu Y. Self-knowledge guided retrieval augmentation for large language models. In *Proc. the 2023 Findings of the Association for Computational Linguistics*, Dec. 2023, pp.10303–10315. DOI: 10.18653/v1/2023.findings-emnlp.691.

[82] Liu Z Y, Lin Y K, Sun M S. Representation Learning for Natural Language Processing. Springer, 2020. DOI: 10.1007/978-981-15-5573-2.

[83] Karpukhin V, Oguz B, Min S, Lewis P, Wu L, Edunov S, Chen D Q, Yih W T. Dense passage retrieval for open-domain question answering. In *Proc. the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Nov. 2020, pp.6769–6781. DOI: 10.18653/v1/2020.emnlp-main.550.

[84] Thulke D, Daheim N, Dugast C, Ney H. Efficient retrieval augmented generation from unstructured knowledge for task-oriented dialog. arXiv: 2102.04643, 2021. https://arxiv.org/abs/2102.04643, May 2024.

[85] Nogueira R, Cho K. Passage re-ranking with BERT. arXiv: 1901.04085, 2019. https://arxiv.org/abs/1901.04085, May 2024.

[86] Peng W J, Li G Y, Jiang Y, Wang Z L, Ou D, Zeng X Y, Xu T, Chen E H. Large language model based long-tail query rewriting in Taobao search. arXiv: 2311.03758, 2023. https://arxiv.org/abs/2311.03758, May 2024.

[87] Gao L Y, Ma X G, Lin J, Callan J. Precise zero-shot dense retrieval without relevance labels. In *Proc. the 61st Annual Meeting of the Association for Computational Linguistics (Volume* 1), Jul. 2023, pp.1762–1777. DOI: 10.18653/v1/2023.acl-long.99.

[88] Zhang W, Feng Y, Meng F D, You D, Liu Q. Bridging the gap between training and inference for neural machine translation. In *Proc. the 57th Annual Meeting of the Association for Computational Linguistics*, Jul. 2019, pp.4334–4343. DOI: 10.18653/v1/P19-1426.

[89] Gu J T, Lu Z D, Li H, Li V O K. Incorporating copying mechanism in sequence-to-sequence learning. In *Proc. the 54th Annual Meeting of the Association for Computational Linguistics (Volume* 1), Aug. 2016, pp.1631–1640. DOI: 10.18653/v1/P16-1154.

[90] See A, Liu P J, Manning C D. Get to the point: Summarization with pointer-generator networks. In *Proc. the 55th Annual Meeting of the Association for Computational Linguistics (Volume* 1), Jul. 2017, pp.1073–1083. DOI: 10.18653/v1/P17-1099.

[91] He S Z, Liu C, Liu K, Zhao J. Generating natural answers by incorporating copying and retrieving mechanisms in sequence-to-sequence learning. In *Proc. the 55th Annual Meeting of the Association for Computational Linguistics (Volume* 1), Jul. 2017, pp.199–208. DOI: 10.18653/v1/P17-1019.

[92] Presti L L, Maggiore G, Marino V, Resciniti R. Mobile instant messaging apps as an opportunity for a conversational approach to marketing: A segmentation study. *Journal of Business & Industrial Marketing*, 2022, 37(7): 1432–1448. DOI: 10.1108/JBIM-02-2020-0121.

[93] McLean G, Osei-Frimpong K. Examining satisfaction with the experience during a live chat service encounter-implications for website providers. *Computers in Human Behavior*, 2017, 76: 494–508. DOI: 10.1016/j.chb.2017.08.005.

[94] Okonkwo C W, Ade-Ibijola A. Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence*, 2021, 2: 100033. DOI: 10.1016/j.caeai.2021.100033.

[95] Pradana A, Sing G O, Kumar Y J. SamBot-intelligent conversational bot for interactive marketing with consumer-centric approach. *International Journal of Computer Information Systems and Industrial Management Applications*, 2014, 6: 265–275.

[96] Cui L, Huang S H, Wei F R, Tan C Q, Duan C Q, Zhou M. SuperAgent: A customer service chatbot for E-commerce websites. In *Proc. the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations*, Jul. 2017, pp.97–102.

[97] Chen S Q, Li C L, Ji F, Zhou W, Chen H Q. Review-driven answer generation for product-related questions in E-commerce. In *Proc. the 12th ACM International Conference on Web Search and Data Mining*, Feb. 2019, pp.411–419. DOI: 10.1145/3289600.3290971.

[98] Xu Z T, Cruz M J, Guevara M *et al*. Retrieval-augmented generation with knowledge graphs for customer service question answering. arXiv: 2404.17723, 2024. https://arxiv.org/abs/2404.17723, May 2024.

[99] Rutten L J F, Blake K D, Greenberg-Worisek A J, Allen S V, Moser R P, Hesse B W. Online health information seeking among us adults: Measuring progress toward a healthy people 2020 objective. *Public Health Reports*, 2019, 134(6): 617–625. DOI: 10.1177/0033354919874074.

[100] Baker A, Perov Y, Middleton K, Baxter J, Mullarkey D, Sangar D, Butt M, DoRosario A, Johri S. A comparison of artificial intelligence and human doctors for the purpose of triage and diagnosis. *Frontiers in Artificial Intelligence*, 2020, 3: 543405. DOI: 10.3389/frai.2020.543405.

[101] Cameron G, Cameron D, Megaw G, Bond R, Mulvenna M, O'Neill S, Armour C, McTear M. Assessing the usability of a chatbot for mental health care. In *Proc. the 2018 International Workshops on Internet Science*, Oct. 2018, pp.121–132. DOI: 10.1007/978-3-030-17705-8_11.

[102] Zakka C, Shad R, Chaurasia A *et al*. Almanac — Retrieval-augmented language models for clinical medicine. *NEJM AI*, 2024, 1(2): aioa2300068. DOI: 10.1056/aioa2300068.

[103] Jin M Y, Yu Q K, Shu D, Zhang C, Fan L Z, Hua W Y, Zhu S Y, Meng Y D, Wang Z T, Du M N, Zhang Y F. Health-LLM: Personalized retrieval-augmented disease prediction system. arXiv: 2402.00746, 2024. https://arxiv.org/abs/2402.00746, May 2024.

[104] Gao Y J, Li R Z, Caskey J R, Dligach D, Miller T, Churpek M M, Afshar M. Leveraging a medical knowledge graph into large language models for diagnosis prediction. arXiv: 2308.14321, 2023. https://arxiv.org/abs/2308.14321, May 2024.

[105] Wu K, Wu E, Cassasola A *et al*. How well do LLMs cite relevant medical references? An evaluation framework and analyses. arXiv: 2402.02008, 2024. https://arxiv.org/abs/2402.02008, May 2024.

[106] Hobert S, von Wolff R M. Say hello to your new automated tutor—A structured literature review on pedagogical conversational agents. In *Proc. the 14th Internationale Tagung Wirtschaftsinformatik*, Feb. 2019, pp.301–314.

[107] Wollny S, Schneider J, Di Mitri D, Weidlich J, Rittberger M, Drachsler H. Are we there yet?—A systematic literature review on chatbots in education. *Frontiers in Artificial Intelligence*, 2021, 4: 654924. DOI: 10.3389/frai.2021.654924.

[108] Settles B, Meeder B. A trainable spaced repetition model for language learning. In *Proc. the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1)*, Aug. 2016, pp.1848–1858. DOI: 10.18653/v1/P16-1174.

[109] Long M H. The role of the linguistic environment in second language acquisition. In *Handbook of Research on Language Acquisition*: *Vol. 2. Second Language Acquisition*, Ritchie W C, Bhatia T K (eds.), Academic Press, 1996, pp.413–468.

[110] Chhibber N, Law E. Using conversational agents to support learning by teaching. arXiv: 1909.13443, 2019. https://arxiv.org/abs/1909.13443, May 2024.

[111] McMahan A. Immersion, engagement, and presence: A method for analyzing 3-D video games. In *The Video Game Theory Reader*, Wolf M J P, Perron B (eds.), Routledge, 2013, pp.89–108.

[112] Mateas M, Stern A. Natural language understanding in façade: Surface-text processing. In *Proc. the 2nd Interna-*

tional Conference on Technologies for Interactive Digital Storytelling and Entertainment, Jun. 2004, pp.3–13. DOI: 10.1007/978-3-540-27797-2_2.

[113] Rastogi A, Zang X X, Sunkara S, Gupta R, Khaitan P. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proc. the 34th AAAI Conference on Artificial Intelligence*, Feb. 2020, pp.8689–8696. DOI: 10.1609/aaai.v34i05.6394.

[114] Dingliwal S, Gao B Y, Agarwal S, Lin C W, Chung T, Hakkani-Tür D Z. Few shot dialogue state tracking using meta-learning. In *Proc. the 16th Conference of the European Chapter of the Association for Computational Linguistics*: *Main Volume*, Apr. 2021, pp.1730–1739. DOI: 10.18653/v1/2021.eacl-main.148.

[115] Chen Z Y, Liu B. Lifelong Machine Learning (2nd edition). Springer, 2018.

[116] Mazumder S, Ma N Z, Liu B. Towards a continuous knowledge learning engine for chatbots. arXiv: 1802.06024, 2018. https://arxiv.org/abs/1802.06024, May 2024.

[117] Tuan Y L, Chen Y N, Lee H Y. DyKgChat: Benchmarking dialogue generation grounding on dynamic knowledge graphs. In *Proc. the 2019 EMNLP-IJCNLP*, Nov. 2019, pp.1855–1865. DOI: 10.18653/v1/D19-1194.

[118] Yurchenko O, Cherednichenko O, Trofimova-Herman A, Kupriianov Y. Towards cross-lingual transfer based on self-learning conversational agent model. In *Proc. the 7th International Conference on Computational Linguistics and Intelligent Systems*, Apr. 2023, pp.194–205.

[119] Korbak T, Elsahar H, Kruszewski G, Dymetman M. Controlling conditional language models without catastrophic forgetting. In *Proc. the 39th International Conference on Machine Learning*, Jul. 2022, pp.11499–11528.

[120] Gibson E, Futrell R, Piantadosi S P, Dautriche I, Mahowald K, Bergen L, Levy R. How efficiency shapes human language. *Trends in Cognitive Sciences*, 2019, 23(5): 389–407. DOI: 10.1016/j.tics.2019.02.003.

[121] Schrimpf M, Blank I A, Tuckute G *et al.* The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences of the United States of America*, 2021, 118(45): e2105646118. DOI: 10.1073/pnas.2105646118.

[122] Feine J, Gnewuch U, Morana S, Maedche A. A taxonomy of social cues for conversational agents. *International Journal of Human-Computer Studies*, 2019, 132: 138–161. DOI: 10.1016/j.ijhcs.2019.07.009.

[123] Xu J, Wang H F, Niu Z Y, Wu H, Che W X, Liu T. Conversational graph grounded policy learning for open-domain conversation generation. In *Proc. the 58th Annual Meeting of the Association for Computational Linguistics*, Jul. 2020, pp.1835–1845. DOI: 10.18653/v1/2020.acl-main.166.

[124] Mirzababaei B, Pammer-Schindler V. Developing a conversational agent's capability to identify structural wrongness in arguments based on Toulmin's model of arguments. *Frontiers in Artificial Intelligence*, 2021, 4: 645516. DOI: 10.3389/FRAI.2021.645516.

[125] Nogueira D M, Maciel C, Viterbo J, Vecchiato D. A privacy-driven data management model for smart personal assistants. In *Proc. the 5th Int. Conf. Human Aspects of Information Security, Privacy and Trust*, Jul. 2017, pp.722–738. DOI: 10.1007/978-3-319-58460-7_49.

[126] Dubiel M, Halvey M, Azzopardi L. A survey investigating usage of virtual personal assistants. arXiv: 1807.04606, 2018. https://arxiv.org/abs/1807.04606, May 2024.

[127] Motger Q, Franch X, Marco J. Software-based dialogue systems: Survey, taxonomy, and challenges. *ACM Computing Surveys*, 2022, 55(5): 91. DOI: 10.1145/3527450.

[128] Lai S T, Leu F Y, Lin J W. A banking chatbot security control procedure for protecting user data security and privacy. In *Proc. the 13th Int. Conf. Broadband and Wireless Computing, Communication and Applications*, Oct. 2018, pp.561–571. DOI: 10.1007/978-3-030-02613-4_50.

[129] Laranjo L, Dunn A G, Tong H L *et al.* Conversational agents in healthcare: A systematic review. *Journal of the American Medical Informatics Association*, 2018, 25(9): 1248–1258. DOI: 10.1093/jamia/ocy072.

[130] Escobar-Planas M. Towards trustworthy conversational agents for children. In *Proc. the 21st Annual ACM Interaction Design and Children Conference*, Jun. 2022, pp.693–695. DOI: 10.1145/3501712.3538826.

[131] Huang X W, Ruan W J, Huang W *et al.* A survey of safety and trustworthiness of large language models through the lens of verification and validation. arXiv: 2305.11391, 2023. https://arxiv.org/abs/2305.11391, May 2024.

Fabio Caffaro is a data scientist. He received his B.S. degree in computer science and his M.S. degree in data science and engineering from Politecnico di Torino, Turin, in 2019 and 2022, respectively. Since 2022, he has been working as an AI applied researcher at LINKS Foundation, Turin, focusing on NLP and time-series forecasting.

Giuseppe Rizzo is a computer scientist with a Ph.D. on natural language processing (NLP). He is a research program manager at LINKS Foundation, Turin, an adjunct professor at Politecnico di Torino, Turin, and a member of the ELLIS network. He has 15 years of experience in research and innovation having worked at the intersection of machine learning and knowledge graph technologies applied to impact areas such as social inclusion, tourism, and health.