# The Role of Summarization in Generative Agents: A Preliminary Perspective

# Xiachong Feng, Xiaocheng Feng, Bing Qin

Harbin Institute of Technology, China

{xiachongfeng,xcfeng,bqin}@ir.hit.edu.cn

### **Abstract**

Generative agents (Park et al., 2023) that simulate human society show tremendous potential for further research and practical applications. Specifically, the generative agent architecture comprising several meticulously designed modules constitutes the most critical component. To facilitate progress in this research, this report presents our integrated perspective on comprehending generative agents through summarization, since we believe summarization is the most fundamental and indispensable capacity of generative agents manifested across diverse scenarios. We hope this report can provide insight into understanding the importance of summarization capacity in generative agents and motivate future research.

# 1 Introduction

Recent advancements in Large Language Models (LLMs), such as ChatGPT and GPT-4 (OpenAI, 2023), have rebuilt various domains including natural language processing (Yang et al., 2023), computer vision (Wu et al., 2023) and autonomous robotics (Mai et al., 2023). These cutting-edge models enable novel opportunities to achieve artificial general intelligence (AGI). Owing to the rapid progress of LLMs, there is an emerging consensus that LLMs have attained preliminary intelligence and now demonstrate comparable performance to humans on various tasks (Zhao et al., 2023).

In the current era of large language models, Park et al. (2023) propose Generative Agents: sophisticated computational software powered by fundamental language models that can simulate believable human behaviour within meticulously designed environments and protocols. This well-designed framework offers comprehensive opportunities for exploring and understanding human social dynamics, including long-term goal planning, information transformation, relationship establishment and coordination.

In this report, we present our view on generative agents from the perspective of automatic summarization and demonstrate how various functional components of such agents can be formalized as summarization tasks. Specifically, we identify several key summarization techniques that are integral to implementing generative agents: (1) The retrieve module contains the idea of unsupervised summarization (§3.1); (2) The reflection module is composed of two sub-modules: extreme summarization (§3.2) and citation-based summarization (§3.3); (3) Query-based summarization (§3.4) supports following Plan module and Act module; (4) Summarization with emojis (§3.5) provides an intuitive visual interface; (5) The agent's movement in the environment can be abstracted to Graph Summarization (§3.6); and (6) Dialogue between agents is facilitated by Dialogue Summarization (§3.7). We hope this paper illuminates the potential of summarization techniques in advancing the development of future generative agents.

# 2 Generative Agents

Generative agents are AI-powered computational software that can simulate believable human behaviour. In this section, we provide a concise overview of the generative agent architecture including several main components. The overall architecture is shown in Figure 1. The fundamental Memory module is responsible for storing various types of information related to the agent itself, including basic observations as well as high-level reflections and generated plans. The **Retrieve** module then extracts appropriate memories from the memory stream to support downstream modules including Plan, Reflect and Act. Afterwards, the **Reflect** module provides high-level abstractions of one agent's memories, which serve as another type of memory. Furthermore, the **Plan** module takes the agent's summary and the observed entity's summary into consideration and creates the plan in a

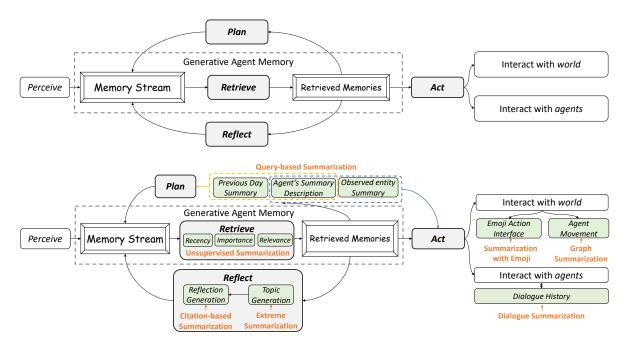


Figure 1: Illustration of the generative agent architecture and key summarization techniques inside the architecture.

course-to-fine manner. Finally, agents **Act** with the world by performing actions or with other agents by initiating dialogues.

# 3 Key Summarization Techniques

#### 3.1 Unsupervised Summarization

The **Retrieve** module aims to offer pilot memories given the agent's current situation and the entire memory stream. This coincides with the objective of unsupervised summarization which seeks to extract the essential information from a collection of documents given one desired query based on various manufactured features (Carbonell and Goldstein-Stewart, 1998).

Specifically, the retrieve function takes three distinct features, Recency, Importance, and Relevance, into consideration to effectively derive prominent information from the memory stream. In detail, Recency posits that recently accessed memories are important since we human beings are frequently processing short-term tasks. Importance is inferred directly by the LLM depending on its tremendous background knowledge by simply prompting the LLM. Relevance assigns a higher score to those most relevant memories with respect to the agent's current situation. The final retrieval score is a weighted sum of three scores: score =  $\alpha_{\text{recency}} \cdot \text{recency} + \alpha_{\text{importance}} \cdot$ importance +  $\alpha_{\text{relevance}}$  · relevance. With the integration of the above three features, the retrieve

module successfully conducts unsupervised summarization over the memory to produce digest information for the following steps.

#### 3.2 Extreme Summarization

**Reflection** is one of the most critical components of the generative agent, which summarizes the agent's recent situation and creates high-level thoughts. The whole reflection can be divided into two steps, in which the first step is extreme summarization. Concretely speaking, it aims to condense the agent's 100 most recent memory records into three key topics in the question generation manner.

Specifically, the module achieves the goal by prompting the LLM via "Given only the information above, what are 3 most salient high-level questions we can answer about the subjects in the statements?". The results include three highly condensed questions that can be viewed as extreme summaries of the agent's recent memories since several previous studies verify the tight connection between summarization and question generation (Narayan et al., 2020; Feng et al., 2021).

#### 3.3 Citation-based Summarization

The second step of **Reflection** can be viewed as a citation-based summarization task, which receives several retrieved documents (memories) with indexes and aims to produce summaries with evidence references. This is also in line with the previous related work generation task and the open-

domain reading comprehension task, both of which require providing concrete evidence to support their generated results (Chen et al., 2021).

Specifically, the module achieves the goal by prompting the LLM via "Statements about Klaus Mueller, ..., What 5 high-level insights can you infer from the above statements? (example format: insight (because of 1, 5, 3))". The output abstracts relevant memories into the reflection with citations: "Klaus Mueller is dedicated to his research on gentrification (because of 1, 2, 8, 15)".

## 3.4 Query-based Summarization

In fact, query-based summarization permeates the core architecture of the entire generative agent with the help of the **Retrieve** module. In this part, we mainly focus on three tasks that will support the subsequent **Plan** and **Act** modules.

Agent's Summary Description Agent's summary description summarizes the agent's identity information, current occupation situation and self-assessment, which serves as a critical clue to making plans and taking reactions. In detail, relevant memories are first obtained via three queries "[name]'s core characteristics", "[name]'s current daily occupation", and "[name's] feeling about his recent progress in life", and then three resulting summaries are combined into the whole agent's summary description.

**Previous Day Summary** Previous Day Summary plays an important role in the plan creation process, which ensures the agent achieves consistent and long-term goals. Although no detailed information is provided in the original paper (Park et al., 2023), we assume the implicit query "[name]'s previous day plan" is used to retrieve relevant memories and produce the final summary.

Observed Entity Summary The observed entity summary that compresses (1) the relationship between the agent and the entity and (2) the status of the entity is an important basis for whether the agent takes action. The summary consists of two parts obtained via queries "What is [observer]'s relationship with the [observed entity]?" and "[Observed entity] is [action status of the observed entity]". Taking both agent's summary description and observed entity summary into consideration, the agent decides whether or not to react by prompting the LLM "Should John react to the observation, and if so, what would be an appropriate reaction?"

## 3.5 Summarization with Emojis

To give quick access to the agent's status, Park et al. (2023) implements a high-level emoji-based abstraction on the sandbox interface by prompting the LLM. For example, "Isabella Rodriguez is checking her emails" appears as ■ . As the saying goes, a picture is worth a thousand words, the emoji interface intuitively summarizes the agent's current status and integrates into the whole system.

#### 3.6 Graph Summarization

Agents who lived in Smallville can perform movements to reach the appropriate location. The Smallville realizes a tree representation, where the root node denotes the entire world, children nodes describe areas and leaf nodes indicate objects. The agent's movement is decided by first transforming the tree representation into natural language and then prompting the LLM via "Which area should [name] go to?". In other words, the movement of an agent can be formalized as an implicit graph summarization task (Kaushik, 2003). Given the world graph, the agent finds one suitable path from the current location towards the target destination.

## 3.7 Dialogue Summarization

The agents interact with each other through dialogue. At the initial point, one agent decides to trigger the dialogue based on the action given the agent's summary description and observed agent (entity) summary. To make the dialogue coherent and informative, the following utterances are generated by considering additional dialogue summaries. In the original paper (Park et al., 2023), pure dialogue histories are used to facilitate dialogue generation. We believe that when facing long and verbose dialogue histories, dialogue summarization can be an effective method to address such a challenge. Additionally, on the demo page, the dialogue summary also provides a quick overview of the core contents of a dialogue.

#### 4 Conclusion

In this report, we aim to understand generative agents from a unified view of summarization. We systematically analyze several key summarization techniques and show how individual modules inside the generative agent architecture can be formalized as traditional summarization tasks. We believe future generative agents can be substantially enhanced with advanced summarization abilities.

#### References

- Jaime G. Carbonell and Jade Goldstein-Stewart. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.
- Xiuying Chen, Hind Alamro, Mingzhe Li, Shen Gao, Xiangliang Zhang, Dongyan Zhao, and Rui Yan. 2021. Capturing relations between scientific papers: An abstractive model for related work section generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6068–6077, Online. Association for Computational Linguistics.
- Xiachong Feng, Xiaocheng Feng, Bing Qin, and Xinwei Geng. 2021. Dialogue discourse-aware graph model and data augmentation for meeting summarization. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3808–3814. International Joint Conferences on Artificial Intelligence Organization. Main Track.
- Shriraghav Kaushik. 2003. *Graph summarization for indexing paths in graph-structured data*. The University of Wisconsin-Madison.
- Jinjie Mai, Jun Chen, Bing chuan Li, Guocheng Qian, Mohamed Elhoseiny, and Bernard Ghanem. 2023. Llm as a robotic brain: Unifying egocentric memory and control. *ArXiv*, abs/2304.09349.
- Shashi Narayan, Gonçalo Simães, Ji Ma, Hannah Craighead, and Ryan T. McDonald. 2020. Qurious: Question generation pretraining for text generation. *ArXiv*, abs/2004.11026.
- OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.
- Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.
- Chenfei Wu, Sheng-Kai Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *ArXiv*, abs/2303.04671.
- Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Z. Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu,

Jianyun Nie, and Ji rong Wen. 2023. A survey of large language models. *ArXiv*, abs/2303.18223.