An Evaluation of Large Language Models for Procedural Action Anticipation

Zeyun Zhong

Vision and Fusion Laboratory
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT), Germany
zeyun.zhong@kit.edu

Abstract

This study evaluates large language models (LLMs) for their effectiveness in long-term action anticipation. Traditional approaches primarily depend on representation learning from extensive video data to understand human activities, a process fraught with challenges due to the intricate nature and variability of these activities. A significant limitation of this method is the difficulty in obtaining effective video representations. Moreover, relying solely on video-based learning can restrict a model's ability to generalize in scenarios involving long-tail classes and out-of-distribution examples. In contrast, the zero-shot or few-shot capabilities of LLMs like ChatGPT offer a novel approach to tackle the complexity of long-term activity understanding without extensive training. We propose three prompting strategies: a plain prompt, a chain-of-thought-based prompt, and an in-context learning prompt. Our experiments on the procedural Breakfast dataset indicate that LLMs can deliver promising results without specific fine-tuning.

1 Introduction

Understanding human activities from video data presents significant challenges due to the inherent variability and complexity of these activities. Traditional methods, which rely heavily on learning representations from large-scale video datasets, face two key limitations. First, the intricacy of human activities makes it difficult to obtain comprehensive representations, especially for longer videos. Second, dependence on extensive datasets restricts the models' ability to generalize to less common, long-tail classes and unseen scenarios.

Recent research has begun to explore the use of large language models [16] (LLMs) to overcome these challenges. These models, equipped with billions of parameters, can utilize training data aggregated from vast, unlabeled text corpora, and have demonstrated exceptional few-shot and zero-shot performance across various tasks. Prior methods [19, 12] have used LLMs for egocentric action anticipation, typically integrating an action recognition model to supply the LLMs with historical action sequence. However, this integration complicates the LLMs' process, leading to less intuitive results.

In this study, we aim to utilize LLMs for long-term action anticipation, minimizing the dependency on action recognition models. We evaluate procedural activities such as breakfast preparation, relying on ground-truth action histories. Our approach departs from traditional methods that are reliant on extensive video data and are limited by the respective training data distributions, focusing instead on the procedural knowledge and generalization ability of LLMs. We design three prompting strategies based on chain-of-thought [18] and in-context learning [1]. These strategies enable LLMs to anticipate future actions by providing a sequence of past observed actions in discrete text. Our experimental results on the Breakfast dataset demonstrate the effectiveness of LLMs in understanding the human activities, showcasing the potential of LLMs in a new domain of activity prediction and understanding.

2 Related Work

Action Anticipation aims to predict future actions given a video clip of the past and present. Many approaches initially investigated different forms of action and activity anticipation from third person video [7, 5, 9]. Recently, along with development of multiple challenge benchmarks [2, 3, 15, 10], the first-person (egocentric) vision has also gained popularity. To accurately predict future actions, the summarization of temporal progression of past actions is essential. To model the past action progression, earlier methods mainly used RNN [5, 6] or TCN [11]-based architectures, which have been shown to be inferior to the recent Transformer-based approaches [8, 21, 9, 22]. Based on the predicted time horizon, action anticipation approaches can be broadly grouped into two categories [20]: short-term anticipation approaches [2, 3] and long-term anticipation approaches [5, 10]. While short-term approaches predict actions a few seconds into the future, long-term approaches aim to predict a sequence of future actions (with their durations) up to several minutes into the future.

Large language models [1, 17] have significantly influenced the natural language processing (NLP) field, exhibiting an impressive capacity to generalize across unseen tasks. With their extensive training data and large parameter size, LLMs have demonstrated the ability to learn from examples provided in input prompts, a concept known as in-context learning [1]. Additionally, LLMs utilize a chain-of-thought [18] reasoning approach. This involves breaking down complex questions into simpler sub-questions, which are then sequentially addressed. This step-by-step reasoning enhances the accuracy and coherence of responses, especially for complex queries, and provides a transparent rationale for the model's thought process.

3 Method

To assess the effectiveness of LLMs in action anticipation, we utilize a procedural dataset, the Breakfast [13] dataset. The LLM is tasked with predicting future actions based on an input sequence of observed human actions, $[a_1, \ldots, a_M]$, where M represents the total number of observed actions. The objective is

Role:

```
{’role’: ’system’,
’content’: ’You are a predictive AI assistant
    focused on Breakfast preparation. All fine-
    grained action classes are: [...].’}
```

Plain Prompt:

```
{’role’: ’user’,
’content’: ’Given the observed fine-grained actions
    : [...], predict the next {N} actions using only
     the predefined action classes. Do not include
    actions outside the predefined list. Respond
    only in this format: <action1>, <action2>, ...’}
```

Figure 3.1: Plain setup. The LLM model is asked to predict N future actions based on a sequence of observations.

to forecast a subsequent series of N actions, $[a_{M+1}, \ldots, a_{M+N}]$, which the human actor is likely to perform. To achieve this goal, we introduce three prompts that are described in the following paragraphs. Additionally, we outline a post-processing methedology to align the outputs of LLMs with the desired format requirements.

Prompt Design. In our initial approach, we present a straightforward setup, as illustrated in Fig. 3.1. To improve the output quality of the LLM model (ChatGPT in our case), we configure the model as a predictive AI assistant specifically tailored for breakfast preparation tasks. To restrict the scope of predictions, we incorporate all action classes from the dataset into the model's setup. This prompt includes the task description and defines the input, i.e., a sequence of observations. In addition, the prompt also defines the output format, mandating the model to predict only actions that are contained in the predefined list.

In our second approach, we adopt a top-down approach [19], utilizing chain-of-thoughts prompts [18] (CoT), as illustrated in Fig. 3.2. This approach initially deduces the overarching activity from the history of actions and then formulates

Role:

```
{'role': 'system',
'content': 'You are a predictive AI assistant
    focused on Breakfast preparation. All fine-
    grained action classes are: [...]. All
    activities are: [...].'}
```

Intention Prompt:

```
{'role': 'user',
'content': 'Given the observed fine-grained actions
    : [...], identify the current activity based on
    these actions and then predict the subsequent {N
    } actions. Use only the predefined action and
    activity classes. Do not include actions or
    activities outside the predefined list. Respond
    only in this format: <activity>; <action1>, <
    action2>, ...'}
```

Figure 3.2: Top-down setup. The LLM model is first asked to identify the current activity given a sequence of observations, and then to predict N future actions based on both the inferred high-level activity and observations.

a plan considering both the historical actions and the intended goal. We construct two CoT questions: `Q1.What's the current activity according to previous actions? Q2.What are the future actions based on the inferred activity and previous actions?` To limit the predictive range for activity forecasts, we also incorporate all high-level activity classes into the model's setup.

In our last approach, we incorporate a few examples from the training set into the prompt to enable in-context learning [1] (ICL), as outlined in Fig. 3.3. Unlike fine-tuning, which involves backward passes through the entire or partial model, ICL leverages the inherent generalization capabilities of LLMs without being constrained to particular datasets or scenarios.

Inference of LLM and Post-processing. It is important to recognize that the outputs generated by LLMs may not always adhere to the required for-

```
'Given the observed fine-grained actions: [...],
    identify the current activity based on these
    actions and then predict the subsequent {N}
    actions. Use only the predefined action and
    activity classes. Do not include actions or
    activities outside the predefined list.
Example 1 - Observed: [...], Activity: [...],
    Predicted actions: [...].
...
Example 4 - Observed: [...], Activity: [...],
    Predicted actions: [...].
Respond only in this format: <activity>; <action1>,
     <action2>, ...'
```

Figure 3.3: Top-down prompt with in-context learning (ICL). A few examples are added to the prompt in Fig. 3.2 to enable in-context learning.

mat and taxonomy, even when the input prompts explicitly include classes from a predefined domain and request predictions within a certain format. For instance, the LLM model might predict `Activity: making tea\nNext predicted action: pour_water`, which deviates from the expected format of `<activity>; <actions>`. This discrepancy complicates the process of metric calculation. To address this, we implement a string matching rule to identify relevant activity or actions for metric evaluation. For simplicity, predictions that fall outside the predefined list are considered false predictions.

4 Experiments

Dataset. The Breakfast [13] dataset comprises 1,712 videos of 52 different individuals making breakfast in 18 different kitchens, totalling 77 hours. Every video is categorized into one of the 10 activities related to breakfast preparation. The videos are annotated by 48 fine-grained actions.

Prompt	Recognition	Anticipation		
	Top-1 ↑	Top-1 ↑	Top-1 agnostic ↑	Edit ↓
Plain	–	12.08±0.90	31.78±0.97	0.87±0.02
Top-down	62.01±2.37	14.83±0.93	33.41±0.90	0.84±0.01
Top-down ICL	94.07±1.14	35.05±1.20	66.46±1.50	0.55±0.02

Table 4.1: Comparison of three presented prompting strategies on the Breakfast [13] dataset. We report the mean performance and the standard deviation of five runs.

Metrics. We evaluate three metrics in this work: top-1 accuracy, top-1 order-agnostic accuracy, and edit distance (ED). Top-1 accuracy measures the precision of predictions in their chronological sequence. In contrast, top-1 order-agnostic accuracy focuses on the presence of correct predictions, regardless of their temporal order, reflecting scenarios where identifying future key actions is crucial, irrespective of preceding or succeeding actions. For instance, in robotic applications, foreseeing a need for assistance, such as an object hand-over, is pivotal, while the exact sequence of preceding or subsequent human actions is less critical. Additionally, we adopt ED [4, 14] to assess the sequential alignment of predictions with actual events. ED incorporates insertions, deletions, substitutions, and transpositions in the predicted actions. A lower ED indicates a higher similarity between the predicted and actual sequences.

Evaluation Details. In our experiments, we utilize the ChatGPT-3.5-turbo [16] model, to serve as the LLM model. The LLM model processes two observed past actions ($M = 2$) and forecasts the next N actions. N is a variable number in our setup and is set as the number of ground-truth actions minus the observed actions for each sequence. In the in-context learning (ICL) setup, illustrated in Fig. 3.3, we select four training examples, following [12]. Specifically, for each sequence in the test set, we identify diverse examples in the training set that include the observed test actions. If the number of identified training examples exceed four, only the initial four are chosen. Conversely, if less than four examples are found, all available examples are utilized in the ICL setup.

Results. In each experimental setup, we execute the Large Language Model (LLM) five times and present both the mean performance and the standard deviation in Table 4.1. Beyond evaluating the accuracy of future action predictions, we also compute the Top-1 accuracy for activity inference in top-down approaches. The results indicate a consistent enhancement in anticipation capabilities through the top-down method, which prioritizes identifying the current activity before predicting future actions. Furthermore, when in-Context learning (ICL) is activated using select examples from the training set, there is a notable improvement in both anticipation and activity recognition performance. Specifically, anticipation accuracy approximately doubles, and activity recognition accuracy increases by 52% (62.01 → 94.07). Additionally, the relatively narrow standard deviation across all metrics suggests that the LLM effectively leverages the provided context to refine its outputs.

5 Conclusion

In this study, we conduct an extensive evaluation of large language models, such as ChatGPT-3.5-turbo, focusing on their capability for long-term action anticipation, particularly leveraging their impressive zero-shot and few-shot learning abilities. This evaluation utilizes the procedural Breakfast dataset. Our findings indicate that these Large Language Models (LLMs) can accurately recognize current activities at an early stage and demonstrate commendable performance in predicting future actions. This underscores the potential of using LLMs for long-term anticipation tasks within the language domain. In the future, we aim to extend the application of LLMs to real-world anticipation scenarios by integrating an action recognition model.

References

[1] Tom Brown et al. “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[2] Dima Damen et al. “Scaling egocentric vision: The epic-kitchens dataset”. In: *ECCV*. 2018, pp. 720–736.

[3] Dima Damen et al. “The epic-kitchens dataset: Collection, challenges and baselines”. In: *TPAMI* 43.11 (2020), pp. 4125–4141.

[4] Fred J Damerau. “A technique for computer detection and correction of spelling errors”. In: *Communications of the ACM* 7.3 (1964), pp. 171–176.

[5] Yazan Abu Farha, Alexander Richard, and Juergen Gall. “When Will You Do What? - Anticipating Temporal Occurrences of Activities”. In: *CVPR*. 2018. arXiv: `1804.00892`.

[6] Antonino Furnari and Giovanni Farinella. “What Would You Expect? Anticipating Egocentric Actions With Rolling-Unrolling LSTMs and Modality Attention”. In: *ICCV*. 2019.

[7] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. “RED: Reinforced Encoder-Decoder Networks for Action Anticipation”. In: *BMVC*. 2017.

[8] Rohit Girdhar and Kristen Grauman. “Anticipative Video Transformer”. In: *ICCV*. 2021. arXiv: `2106.02036`.

[9] Dayoung Gong et al. “Future Transformer for Long-term Action Anticipation”. In: *CVPR*. 2022. arXiv: `2205.14022 [cs]`.

[10] Kristen Grauman et al. “Ego4D: Around the World in 3,000 Hours of Egocentric Video”. In: *CVPR*. 2022. arXiv: `2110.07058`. (Visited on 04/28/2022).

[11] Qiuhong Ke, Mario Fritz, and Bernt Schiele. “Time-Conditioned Action Anticipation in One Shot”. In: *CVPR*. June 2019.

[12] Sanghwan Kim et al. “LALM: Long-Term Action Anticipation with Language Models”. In: *arXiv preprint arXiv:2311.17944* (2023).

[13] Hilde Kuehne, Ali Arslan, and Thomas Serre. “The Language of Actions: Recovering the Syntax and Semantics of Goal-Directed Human Activities”. In: *CVPR*. 2014.

[14] Vladimir I Levenshtein et al. “Binary codes capable of correcting deletions, insertions, and reversals”. In: *Soviet physics doklady*. Vol. 10. 8. 1966, pp. 707–710.

[15] Yin Li, Miao Liu, and James M Rehg. “In the eye of beholder: Joint learning of gaze and actions in first person video”. In: *ECCV*. 2018, pp. 619–635.

[16] OpenAI. *Chatgpt: Optimizing language models for dialogue*. 2022.

[17] Hugo Touvron et al. “Llama: Open and efficient foundation language models”. In: *arXiv preprint arXiv:2302.13971* (2023).

[18] Jason Wei et al. “Chain-of-thought prompting elicits reasoning in large language models”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 24824–24837.

[19] Qi Zhao et al. “AntGPT: Can Large Language Models Help Long-term Action Anticipation from Videos?” In: *arXiv preprint arXiv:2307.16368* (2023).

[20] Zeyun Zhong et al. “A Survey on Deep Learning Techniques for Action Anticipation”. In: *arXiv preprint arXiv:2309.17257* (2023).

[21] Zeyun Zhong et al. “Anticipative Feature Fusion Transformer for Multi-Modal Action Anticipation”. In: *WACV*. 2023.

[22] Zeyun Zhong et al. “DiffAnt: Diffusion Models for Action Anticipation”. In: *arXiv preprint arXiv:2311.15991* (2023).