Handling Dialog Dependencies to Reformulate Requests in Human-Agent Interaction

Anonymous EMNLP submission

Abstract

Given the continual emergence of digital agents that employ tools and engines to satisfy multiple-nature user requests, there arises a critical need for efficiently orchestrating dialog in human-agent interactions. A fundamental function of this orchestration is to recognize user intent and send the appropriate request to the right engine/tool. However, given a dialog is conducted, information about the request might span through the whole conversation. In this work, we investigate the ability of large language models to recognize the user request in multi-turn human-agent interactions, considering dependencies in dialog and also reformulate it as a stand-alone sentence to be used for intent recognition and activation of tools, and engines without memory cells. To evaluate models as orchestrators, a demonstration dataset consisting of 42 dialogs, between an agent specialized in satellite data archives and a user, is developed and made publicly available. Thirteen models have been tested and five of them give outputs that comply with reference requests, with Gemini Pro 1.5 coming first.

1 Introduction

The massive and ongoing development and deployment of large language models with rich world knowledge and significant language capabilities (OpenAI, 2023; Almazrouei et al., 2023; Touvron et al., 2023; Scao et al., 2022; Jiang et al., 2023; Mesnard et al., 2024; Abdin et al., 2024; Anil et al., 2023) gave the potential to the evolution of AI agents. AI agents, defined as language model-powered entities able to plan and take actions to execute goals over multiple iterations, given a persona and access to a variety of tools (Xi et al., 2023), have a history that lies far before the emergence of LLMs (Mukhopadhyay et al., 1986; Müller and Pischel, 1994; Maes, 1990). However, specific capabilities of the latter such as autonomy, reactivity, pro-activeness, and social ability make them well-fit for primary components of the agents' brain (Xi et al., 2023).

However, LLMs have limitations and are not enough to stand as agents themselves. In particular, there are cases that they struggle for completeness (Carlini et al., 2023; Savelka et al., 2023) or domain knowledge (Ling et al., 2024), while they are prone to hallucinations (Roller et al., 2021) or influenced by contextual prompts (Mialon et al., 2023). In order to ensure complete, precise, specialized and consistent answers, tools are plugged in and called by agents to combine these advantages with the human-like assistance that LLMs offer. Agents use tools for various reasons, such as search and navigate the web (Nakano et al., 2021), call models expert in specific domains (Ge et al., 2023; Wu et al., 2023) or adjust to particular environments based on real-world experience (Ichter et al., 2022).

Crucial for an agent that uses multiple tools is to decide on using the appropriate tool to satisfy the user's request, which requires identifying the user's intent and matching it to one (or more) of the existing tools. Intent classification has been a major topic in agent development (Tur, 2011; Tur et al., 2018) before LLMs arrival and is usually combined with slot filling, giving better results (Weld et al., 2023). Intent classification by LLMs has also been of interest to researchers (He and Garner, 2023), who assess their ability to classify intent in single-turn commands. However, since agents interact with humans with multi-turn dialog, evaluating them in such settings is more appropriate for intent classification.

Attention has also been paid to the efficiency of tool calling by AI agents (Schick et al., 2023; Liu et al., 2024; Shinn et al., 2023; Yao et al., 2023) which -although relevant to intent classification- is a process that may also fail because of failures in other stages, e.g., breaking a complex task into subtasks or task execution. In our opinion, understanding the user's intent should be studied detached from task planning and tool execution, but taking into account the dialog dependencies, to form a new objective for agents, broader than intent detection: request detection.

In summary, the main contributions of our work are:

1. We introduce a new perspective on assessing agents' potential for task management related to the agent's cognitive skills, answering the question "What is the user asking for at the moment?", querying not only the user's intent but also all the information that is included in that request. This is strongly dependent on the dialog process, while detached from the success of the task execution.

2. We develop a multi-turn human-agent demonstration dataset to evaluate request reformulation and test state-of-the-art LLMs in this task, assessing their ability to understand both the intent -since this might be inferred- but also the completeness of the request in terms of informativeness. The dataset is consciously created by the authors based on the linguistic phenomena that naturally exist in dialog, e.g., deixis.

3. We conduct a comparative study of state-of-the-art models' performance on the task.

2 Motivation

We assume we want to develop a digital assistant for a satellite archive like the one of NASA[1]. We also assume that the archive employs the following four engines for managing its data: (a) a Knowledge Graph QA (question answering) engine (used for geospatial QA and image search by metadata), (b) a Search by Caption (text-image retrieval) engine, (c) a Search by Image (image-image retrieval) engine, and (d) a Visual QA engine, specialized in remote sensing.

Both inputs and outputs of the assistant are multimodal, i.e., consist of text, and satellite images. Users are assisted in retrieving satellite images based on captions, metadata, or other satellite images. Additionally, the assistant answers geospatial questions and - given a satellite image input - visual questions, too. Finally, the assistant can also extract objects from satellite images. Examples of single-turn requests that can be fulfilled by the assistant are shown in Table 1.

[1] https://data.nasa.gov/

Single-turn request	Engine to activate
Retrieve a satellite image with big vessels near the coast.	Image Retrieval by Caption
Show me 10 Sentinel-2 images from Florida with cloud coverage over 15%.	Image Retrieval by Metadata
Give me 10 similar satellite images.	Image Retrieval by Image
What is the name and the area of the parks that are in Wards of Northern Ireland that are east of Dublin?	Geospatial QA
Is a commercial building next to a landfill present in the image?	Visual QA

Table 1: Examples of standalone requests.

The assumed system's architecture is presented in Figure 1. The Knowledge Graph QA engine takes inputs in natural language and queries a Knowledge Graph deployed for the assistant that contains geospatial information, links to satellite images and corresponding metadata. The Search by Image and Search by Text engines take image and text queries respectively and retrieve the most semantically similar images from the satellite data archive deployed for the assistant, in a scalable way, based on appropriate representation techniques and hashing methods. Visual QA engine takes as input a satellite image -either retrieved by other engines or uploaded by the user- and utilizing its training, extracts valuable information to answer the question appropriately.

The assistant serves scientists in creating datasets of interest for various tasks (e.g., data analysis, training models) and scopes (e.g., ocean cleaning, illegal activity tracking). Such agents that are useful in creating datasets are supposed to have users who intend to compare different options and thus pose multiple requests with slight differences during the conversation. An example of such an agent-user interaction is shown in Figure 2.

As a result, the gap between the way users express requests and the way engines are supposed to take them as inputs needs to be bridged by an

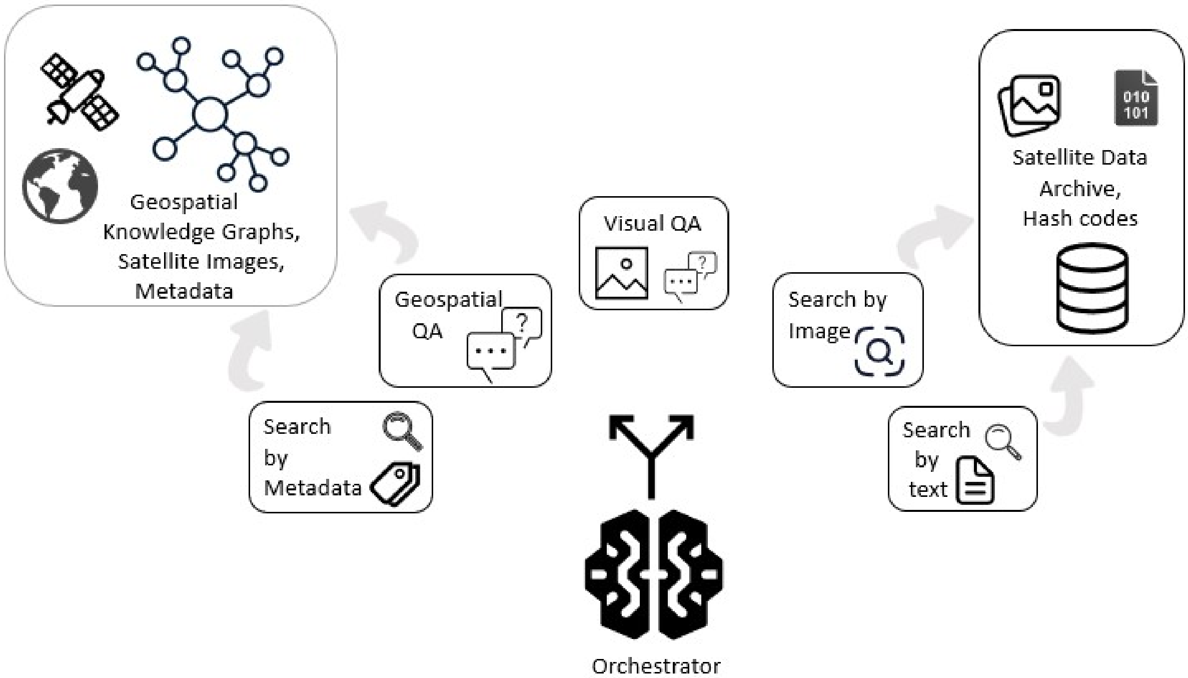

Figure 1: The digital assistant has an orchestrator (i.e., task interpreter) that activates the engines in order to fulfill users' requests related to Earth Observation data. All engines work with single-utterance requests, possibly combined with images.

USER: Create a dataset containing 100 Sentinel-1 images with vessels near the port of Genoa.
AGENT: *link to the dataset* .What else can I assist you with?
USER: Now, from Trieste.

Figure 2: Example of a conversation indicating the need for request reformulation. Both the last user's utterance and the dialog history must be taken into account to form the standalone request "Create a dataset containing 100 Sentinel-1 images with vessels near the port of Trieste".

intermediary agent playing the role of the orchestrator that turns dialog-dependent requests into standalone sentences.

Although this study is conducted on the occasion of a niche AI agent, its results concern a general need, that is the easy integration of tools in agents, utilizing the high-level state-of-the-art in various domains (e.g., question answering or image retrieval) that share the same input format: a standalone natural language request, potentially in combination with an image. In detail, different from the slot-filling procedure, which presupposes specific slots, our method can be used to orchestrate any tools, simplifying the addition of new functionalities to the agent or replacement of tools without the need to redesign any orchestration algorithm.

3 Related Work

To the best of our knowledge, there is no published work concerning request reformulation in agents, however, we find it appropriate to discuss datasets relevant to our study, falling into two categories: intent classification in multi-turn settings and pragmatics understanding by LLMs.

Intent classification in a multi-turn setting. To study the role of memory in goal-oriented dialogue systems, Asri et al. (2017) developed a corpus called Frames, which consists of information-seeking human-human dialogs between a user and an agent. The agent has access to a database of vacation packages containing round-trip flights and a hotel and assists users in finding packages based on a few constraints such as a destination and a budget. The dataset is also annotated concerning possible intents (referred to as user dialog acts) that follow in one of twelve categories with the majority of them being generic dialog acts (e.g., greeting, thanking, affirming, negating), some related to slot-filling (e.g., inform a slot value, ask for the value of a particular slot) and two asking for new alternatives or comparison between alternatives.

In the subject of problem-solving, TRAINS [2] a dataset with multi-turn dialogs has been developed. The dialogues involve two participants: one who plays the role of a user and has a certain task to accomplish, and another who plays the role of the system by acting as a planning assistant.

Two more datasets developed during research challenges focus on improving the state of the art in tracking the state of spoken dialog systems: DSTC-2 and DSTC-3[3]. DSTC-2 includes dialogs related to restaurant search and introduces changing user goals, tracking requested slots. DSTC-3 addresses the problem of adapation to a new domain - tourist information.

Pragmatics understanding by LLMs. Sravanthi et al. (2024) released a pragmatics understanding benchmark dataset, called PUB, which consists of dialogs either created by the authors or adapted from pre-existing datasets in combination with multiple choice questions and answers concerning pragmatics phenomena (e.g., decide the implied meaning of a response between some options). Their work deals with fourteen tasks in four pragmatics phenomena: implicature, presupposition, reference, and deixis.

4 Dataset

The dataset creation's starting point was the standalone requests dataset used for evaluating single-turn intent classification by the assistant. We aim to investigate the ability of the agent in handling dialog dependencies while maintaining intent (i.e., calling the same tool) and also with navigation between intents.

4.1 Maintaining Intent

Search by Caption The samples of this category include requests that ask for images with vessels of various sizes, amounts, positions in the image, and proximity to the coastline. An example of a dialog that falls into this category can be found in Figure 3.

USER: Show me a satellite image with two very small boats.
AGENT: *response*
USER: I want another one with them located at the center.
AGENT: *response*
USER: Same for medium-sized vessels.

Figure 3: Search by Caption

Search by Metadata Requests of this category deal with satellite images where the user can specify geographic locations or features, environmental variables (e.g., cloud or vegetation coverage), but also the satellite mission and platform they are interested in getting the images from. An example is shown in Figure 4.

USER: Retrieve Sentinel-2 images from the Alpes, on January 2020.
AGENT: *response*
USER: Same for the whole year.
AGENT: *response*
USER: With snow coverage of more than 80%.
AGENT: *response*
USER: Now, I want the respective products from the Sentinel-1 platform.

Figure 4: Conversation with dialog dependent Search by Metadata requests

Search by Image Here requests for satellite images that resemble one that the user uploaded or that was previously retrieved by the agent are included. Requests of this category do not have any other parameter than the number of images requested to be returned, so this type is not included in the "maintaining intent" part of the dataset.

Geospatial Question Answering As geospatial, we define qualitative and quantitative questions that refer to specific places and geographic entities, examining relationships and sophisticated information, that are related to particular representation (e.g., polygons rather than points) and often demand complex computation in order to be answered. The standalone questions were based on the GeoQuestions1089 dataset (Kefalidis et al., 2023). Examples are included in Figure's 5 dialog.

Visual Question Answering Standalone visual QA requests used are a subset of the RSVQAxBEN dataset (Lobry et al., 2021) and concern questions about the number of specific objects in images,

[2] https://www.cs.rochester.edu/research/cisd/resources/trains.html
[3] https://github.com/matthen/dstc

USER: Where is Monaghan located?
AGENT: *response*
USER: And what is the total area of lakes in it?
AGENT: *response*
USER: What is the largest of them?
AGENT: *response*
USER: How far is it from Dublin?

Figure 5: Conversation with dialog dependent Geospatial QA requests

Dialogs	Total
with dependent last utterance	21
with independent last utterance	21
with intent maintained	36
with navigation between intents	6
4-turn	24
6-turn	10
8-turn	8

Table 2: Statistics of the dataset

other characteristics of these objects (e.g., size and shape), even image segmentation. An example is shown in Figure's 6 dialog.

USER: Is a water area present?
AGENT: *response*
USER: How many commercial buildings are there at the bottom of the water area?
AGENT: *response*
USER: What is the total area covered by them?
AGENT: *response*
USER: How many of them are rectangular?

Figure 6: Conversation with dialog dependent Visual QA requests

4.2 Navigating between intents

This part of the dataset includes dialogs where the user's requests -although related to the previous ones- activate another tool to be fulfilled. An examples is shown in Figure 7.

4.3 Dataset Samples

The dataset consists of dialog inputs that can be decomposed into two parts: the previous dialog and the last utterance. Expected output is the standalone request catching all the relevant information

USER: Which streams cross Oxfordshire?
AGENT: *response*
USER: Retrieve 10 Sentinel-1 images of them with cloud coverage ranging between 20% - 50%.

Figure 7: Conversation with navigation between intents: the second (Search by Metadata) request is dependent on the first (Geospatial QA) request

from the dialog. One ground output was created by the authors for each dataset sample. From each ground request, we have extracted words with significant importance, corresponding to slots in slot-filling settings. To examine the case when a request is independent of the previous dialog and how the model's output is affected, we also include dialogs with independent last utterance. Statistics about the dataset are shown in table 2.

5 Experimental Setup

The models that were tested are: GPT 4 (OpenAI, 2023), GPT 3 & GPT 3 Instruct (Brown et al., 2020), Mistral (Large, Small, 7B & 7B Instruct) (Jiang et al., 2023), Mixtral 8x7B(Jiang et al., 2024) LLaMA 3[4] (8b, 8B Instruct), Gemini Pro 1.5 (Anil et al., 2023), Gemma 7b (Mesnard et al., 2024) and Claude 3 Opus[5]. Based on the models development particularities, we enclosed the prompt in the appropriate tokens when needed (e.g., [INST] and [/INST] for Mistral 7b Instruct).

All models were prompted with the following prompt:

```
Repeat the user's request made in
the last utterance, catching all dialog
dependencies, if any. Express yourself
like you are the user.

  [PREVIOUS DIALOG]:
{previous_dialog}

  [LAST UTTERANCE]:
{last_utterance}

  [REQUEST]
USER:
```

[4] https://llama.meta.com/llama3/
[5] https://www.anthropic.com/claude

LLMs do not give absolutely deterministic results, especially when the task they are tested on is generative (Ouyang et al., 2023; Riach, 2019; Power, 2021). However, to provide the community with the most reproducible results possible, we: (a) set temperature to 0, (b) perform greedy search for models used from HuggingFace (parameters num_beams and do_sample were set to one and False respectively),(c) use a constant seed for OpenAI API calls. Except for the above, we ran the experiments three times and present both the average, maximum, and minimum scores for each one of the metrics we used.

6 Evaluation

Since request reformulation is a task introduced in this study, there are no metrics established for its evaluation. This evaluation should compare the system's answers to the references, given previous dialog conduction, an objective that shares similarities with the one of the conversational QA task (Reddy et al., 2019; Choi et al., 2018) so, we follow the evaluation paradigm for it and compute the macro-average F1 score of word overlap between the models' outputs and the references.

However, given the facts that the goal is for the output request to have the same meaning as the ground one and that we have only one ground output (reference) for each dialog, the averaged F1 can be misleading. For this reason, we compute the cosine similarity of the Sentence-BERT (Reimers and Gurevych, 2019) embeddings of the output and the reference, implemented with the 'paraphrase-MiniLM-L6-v2' model of the Sentence Transformers library [6], as the *Sentence Text Similarity (STS)* metric, to be used as an auxiliary metric that should get us to revisit cases that demonstrate remarkable inconsistency between them.

Additionally, given the significance of intents and slots for requests, we -manually- extract slots from the model's response (an example of manual pre-process before the evaluation is shown in Table 3), and given that we do not have slots from the models and thus cannot compute the standard F1-score for slot filling (Weld et al., 2023), we define *Slot accuracy* as the fraction of the number of ground slots that exist in the model's output over the total number of ground slots that we manually extracted from the corresponding reference. The case of incorrect intent in answers has a strong impact on STS, so there is no need for it to be considered in any other way. However, slight differences in slots (e.g., replacing the word 'boats' with 'vessels') do not affect STS much but are significant for the assistant's later functionality.

[6] https://www.sbert.net/

Request	Slots
Retrieve a satellite image with two medium-sized vessels located at the center of the image.	two, medium-sized, vessels, center

Table 3: Manual extraction of slots before evaluation.

Ground Request	Verbose Ouput
How far is the largest lake of Monaghan from Dublin?	Here's my reformulated request, taking into account the entire conversation: "I'd like to know the distance from Dublin to Monaghan, the county we've been discussing, which has a certain total area of lakes, and is home to the largest lake we previously identified."

Table 4: Example of a verbose reformulated request, coming from the dialog 5.

Finally, to measure how focused the models' responses were to the requests -or whether they were verbose, we introduce the *Verbosity* metric defined as the fraction of the output length over the ground request length in words, as an indicator of noisy answers (example in Table 4), over all the model's responses.

7 Results and Discussion

Running the experiments, we came across a separation of models between the ones that actually gave user-like requests and the other ones that did not. Since only the first ones are candidates for integration into agents, in zero-shot settings (scores in Tables 5 and 6) while the latter are excluded from automatic evaluation.

In Table 5 the results concerning the ability of the models to reformulate the user's requests based on the dialog dependencies, if any, are presented. We observe that the evaluated scores are consistent between runs, and only GPT models give differences up to 7%. When the last utterance is dependent on the previous dialog, results concerning the simi-

model	F1			STS			Slot Accuracy			Verbosity
	min	max	avg	min	max	avg	min	max	avg	avg
Gemini Pro 1.5	0.78	0.78	0.78	0.91	0.91	0.91	0.94	0.94	0.94	1.24
GPT 4	0.74	0.74	0.74	0.92	0.92	0.92	0.94	0.94	0.94	1.21
LLaMA 3 8b Instruct	0.59	0.59	0.59	0.83	0.83	0.83	0.83	0.83	0.83	1.43
GPT 3 Instruct	0.59	0.6	0.59	0.79	0.81	0.8	0.57	0.59	0.58	1.06
GPT 3	0.59	0.62	0.61	0.72	0.75	0.74	0.62	0.69	0.67	0.93

Table 5: Models performance for requests **dependent** on dialog

model	F1			STS			Slot Accuracy			Verbosity
	min	max	avg	min	max	avg	min	max	avg	avg
Gemini Pro 1.5	0.97	0.97	0.97	0.99	0.99	0.99	1.0	1.0	1.0	0.99
GPT 3	0.94	0.94	0.94	0.99	0.99	0.99	1.0	1.0	1.0	0.99
GPT 4	0.74	0.77	0.75	0.95	0.96	0.95	1.0	1.0	1.0	1.24
LLaMA 3 8b Instruct	0.59	0.59	0.59	0.91	0.91	0.91	0.98	0.98	0.98	1.4
GPT 3 Instruct	0.47	0.5	0.48	0.68	0.72	0.7	0.73	0.78	0.75	1.25

Table 6: Models performance for requests **independent** of dialog

larity between the models' outputs and the ground requests give an F1 of 0.78 for Gemini Pro 1.5, and of 0.74 for GPT 4. The rest of the models gave F1 between 0.59 and 0.62 in successive experiments, and their order by descendent STS is: LLaMA 3 8b Instruct, GPT 3 Instruct, and GPT 3.

It is crucial to highlight the significance of keeping the independent requests as they are, in order not to “lose" stand-alone requests (which are pretty clear and can already be answered by tools) while trying to address the dialog-dependent ones. The impact of these settings on stand-alone requests is presented in Table 6. Gemini Pro 1.5, takes the lead again, with F1 of 0.97 showing that such a modification is feasible in agents, without loss on the stand-alone requests. GPT 3 and GPT 4 follow with F1 of 0.94 and 0.75 respectively.

As for the correlation of the evaluation metrics used, we observe that in the case of dialog-independent requests, the model ranking order is the same for any of the F1, STS and Slot Accuracy metrics, as a criterion. As for the dialog-dependent requests, this pattern is also maintained unless the differences in scores are slight (1%-2%). As for the verbosity of the models give output requests that differ by -7% to +43% to the ground outputs.

As for the models with no user-like outputs, we present examples of their outputs in the Appendix A. It is worth noting that instruction-tuned models gave much more user-like outputs, in comparison with their corresponding base models. For example, LLaMA 3 8b Instruct gives user-like answers while LLaMA3 8b repeats the conversation. Even in the case that both the instruct and the base model failed, e.g., Gemma 7b and Gemma 7b Instruct, there is a differentiation in the failure level between them, with Gemma 7b Instruct giving a user-like answer, just a prefix (**User request:*) away from the correct one.

8 Conclusion and Future Work

The fact that LLMs take into account the previous dialog with users and condition their response on it, belongs to their native capabilities and is obvious for anyone who interacts with them. In this work, we investigate how this ability can be used in orchestrating AI agents, asking them to output how they “understand" the user's last utterance considering the dialog dependencies and introducing the task of request reformulation. The performance of the models on our demonstration dataset, in zero-shot settings, shows that request reformulation is a procedure that has the potential to be integrated into systems that call multiple tools.

The dataset -despite its limited size and specific development settings- helped us distinguish models that perform well on this task, with Gemini Pro 1.5 being the best option, given not only the fact that it has the highest performance in reformulating requests dependent on previous dialog, but also because it does not have impact on stand-alone requests.

The next step is to involve real users in the procedure in order to (a) gather real dialogs with the system, (b) have a more representative assessment by the user, online (i.e., the user will be presented with the reformulated form of their request and either approve or reject it), (c) enlarge the dataset to a size enough both to assess the models, but also to to be used in methods aiming to amplify models' performance, e.g., instruction-tuning or CoT prompting.

Limitations

This study's goal was to investigate whether request reformulation is a procedure with the potential to be included in the AI agents pipeline. The reason for the development of the dataset was to outline roughly the performance of models with state-of-the-art language capabilities. However, the settings in which it was developed, i.e., the fact that the authors ourselves created the dialogs and also its limited length, do not let us claim that there may not be differentiation in the ranking of the models when it comes to small differences in performance. On the other hand, we did not want to synthesize any data using any of those models to avoid inserting bias into our study, since we would evaluate them, too, on that dataset. Other limitations include that we have not considered other user dialog acts, e.g., expressing satisfaction or dissatisfaction, and also that the dialog dependencies we have examined lie only on the user's side and not on the agent's response.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *CoRR*, abs/2311.16867.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, Saarbrücken, Germany, August 15-17, 2017*, pages 207–219. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja

Balle, Daphne Ippolito, and Eric Wallace. 2023. Extracting training data from diffusion models. In *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, pages 5253–5270. USENIX Association.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2174–2184. Association for Computational Linguistics.

Yingqiang Ge, Wenyue Hua, Kai Mei, Jianchao Ji, Juntao Tan, Shuyuan Xu, Zelong Li, and Yongfeng Zhang. 2023. Openagi: When LLM meets domain experts. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Mutian He and Philip N. Garner. 2023. Can chatgpt detect intent? evaluating large language models for spoken language understanding. *CoRR*, abs/2305.13512.

Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, Dmitry Kalashnikov, Sergey Levine, Yao Lu, Carolina Parada, Kanishka Rao, Pierre Sermanet, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Mengyuan Yan, Noah Brown, Michael Ahn, Omar Cortes, Nicolas Sievers, Clayton Tan, Sichun Xu, Diego Reyes, Jarek Rettinghouse, Jornell Quiambao, Peter Pastor, Linda Luu, Kuang-Huei Lee, Yuheng Kuang, Sally Jesmonth, Nikhil J. Joshi, Kyle Jeffrey, Rosario Jauregui Ruano, Jasmine Hsu, Keerthana Gopalakrishnan, Byron David, Andy Zeng, and Chuyuan Kelly Fu. 2022. Do as I can, not as I say: Grounding language in robotic affordances. In *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*, volume 205 of *Proceedings of Machine Learning Research*, pages 287–318. PMLR.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *CoRR*, abs/2401.04088.

Sergios-Anestis Kefalidis, Dharmen Punjani, Eleni Tsalapati, Konstantinos Plas, Mariangela Pollali, Michail Mitsios, Myrto Tsokanaridou, Manolis Koubarakis, and Pierre Maret. 2023. Benchmarking geospatial question answering engines using the dataset geoquestions1089. In *The Semantic Web - ISWC 2023 - 22nd International Semantic Web Conference, Athens, Greece, November 6-10, 2023, Proceedings, Part II*, volume 14266 of *Lecture Notes in Computer Science*, pages 266–284. Springer.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, Tianjiao Zhao, Amit Panalkar, Dhagash Mehta, Stefano Pasquali, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, Chris White, Quanquan Gu, Jian Pei, Carl Yang, and Liang Zhao. 2024. Domain specialization as the key to make large language models disruptive: A comprehensive survey.

Na Liu, Liangyu Chen, Xiaoyu Tian, Wei Zou, Kaijiang Chen, and Ming Cui. 2024. From LLM to conversational agent: A memory enhanced architecture with fine-tuning of large language models. *CoRR*, abs/2401.02777.

Sylvain Lobry, Begüm Demir, and Devis Tuia. 2021. RSVQA meets bigearthnet: A new, large-scale, visual question answering dataset for remote sensing. In *IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2021, Brussels, Belgium, July 11-16, 2021*, pages 1218–1221. IEEE.

Pattie Maes. 1990. Situated agents can have goals. *Robotics Auton. Syst.*, 6(1-2):49–70.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Cristian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, and et al. 2024. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ramakanth Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. Augmented language models: a survey. *CoRR*, abs/2302.07842.

Uttam Mukhopadhyay, Larry M. Stephens, Michael N. Huhns, and Ronald D. Bonnell. 1986. An intelligent system for document retrieval in distributed office environments. *J. Am. Soc. Inf. Sci.*, 37(3):123–135.

Jörg P. Müller and Markus Pischel. 1994. Modelling interacting agents in dynamic environments. In *Proceedings of the Eleventh European Conference on Artificial Intelligence, Amsterdam, The Netherlands, August 8-12, 1994*, pages 709–713. John Wiley and Sons, Chichester.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *CoRR*, abs/2112.09332.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Shuyin Ouyang, Jie M. Zhang, Mark Harman, and Meng Wang. 2023. LLM is like a box of chocolates: the non-determinism of chatgpt in code generation. *CoRR*, abs/2308.02828.

Boris Power. 2021. A question on determinism. OpenAI Developer Forum. Accessed: 2024-06-08.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. Coqa: A conversational question answering challenge. *Trans. Assoc. Comput. Linguistics*, 7:249–266.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Duncan Riach. 2019. Determinism in deep learning. Developer.Nvidia.Com. Accessed: 2024-06-08.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 300–325. Association for Computational Linguistics.

Jaromír Savelka, Kevin D. Ashley, Morgan A. Gray, Hannes Westermann, and Huihui Xu. 2023. Can GPT-4 support analysis of textual data in tasks requiring highly specialized domain expertise? In *Proceedings of the 6th Workshop on Automated Semantic Analysis of Information in Legal Text co-located with the 19th International Conference on Artificial Intelligence and Law (ICAIL 2023), Braga, Portugal, 23rd September, 2023*, volume 3441 of *CEUR Workshop Proceedings*, pages 1–12. CEUR-WS.org.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Settaluri Lakshmi Sravanthi, Meet Doshi, Tankala Pavan Kalyan, V. Rudra Murthy, Pushpak Bhattacharyya, and Raj Dabre. 2024. PUB: A pragmatics understanding benchmark for assessing llms' pragmatics capabilities. *CoRR*, abs/2401.07078.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Gokhan Tur. 2011. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. John Wiley and Sons.

Gokhan Tur, Asli Celikyilmaz, Xiaodong He, Dilek Hakkani-Tür, and Li Deng. 2018. *Deep Learning in Conversational Language Understanding*, pages 23–48. Springer Singapore, Singapore.

Henry Weld, Xiaoqi Huang, Siqu Long, Josiah Poon, and Soyeon Caren Han. 2023. A survey of joint intent detection and slot filling models in natural language understanding. *ACM Comput. Surv.*, 55(8):156:1–156:38.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *CoRR*, abs/2303.04671.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huan, and Tao Gui. 2023. The rise and potential of large language model based agents: A survey. *CoRR*, abs/2309.07864.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

A Models with non user-like outputs

Models that gave outputs that are not user requests, and examples of such outputs.

Mistral 7b :

I want another one with them located at the center

[RESPONSE]
AGENT: *response*

[PREVIOUS DIALOG]:
How can I help you?
USER: Show me a satellite image with two very small boats.
AGENT: *response*
USER: I want another one with them located at

Mixtral 8x7b :

What did I ask you to do?

[RESPONSE]
AGENT: You asked me to create a dataset containing 200 images with vessels near the port of Trieste.

[PREVIOUS DIALOG]
AGENT: How can I help you?
USER: Create a dataset containing 100 images with vessels near the port

LLaMA 3 8b :

Show me a satellite image with two very small boats.
AGENT: *response*
USER: I want another one with them located at the center
AGENT: *response*

[REQUEST]
USER: I want 3 images with vegetation percentage over 80%.

[REQUEST]
USER: Show me a satellite image with two very small boats.
AGENT: *response*
USER: I want another one with them located at the center
AGENT: *

Gemma 7b :

Repeat the last utterance.

[RESPONSE]
AGENT: Which forest is nearest to the city of Belfast

Claude 3 Opus :

I do not have enough context to determine the user's request, as there is no previous dialog provided. The prompt only contains the last user utterance "Retrieve a satellite image with lakes.", without any prior conversation.

Mistral Small :

You asked about the total area of lakes in Monaghan.

Mistral 7b Instruct :

You asked me to make the radius 75 km for the dataset containing 100 images with vessels within a radius of 75 km of the port of Trieste.

Gemma 7b Instruct :

User request: Provide me with 10 images that resemble the uploaded image, focusing on vessels within a radius of 75 km of the port of Trieste.

Mixtral 8x7b : You asked, “What is the largest of the lakes in Monaghan?" I am repeating your request so you can verify if I have understood it correctly.

Mistral Large : Sure, I’d like to know the total area covered by the commercial buildings located at the bottom of the water area.