Referential ambiguity and clarification requests: comparing human and LLM behaviour

Chris Madge, Matthew Purver and Massimo Poesio
Queen Mary University of London
{c.j.madge,m.purver,m.poesio}@qmul.ac.uk

Abstract

In this work we examine LLMs' ability to ask clarification questions in task-oriented dialogues that follow the asynchronous instruction-giver/instruction-follower format. We present a new corpus that combines two existing annotations of the Minecraft Dialogue Corpus — one for reference and ambiguity in reference, and one for SDRT including clarifications — into a single common format providing the necessary information to experiment with clarifications and their relation to ambiguity. With this corpus we compare LLM actions with original human-generated clarification questions, examining how both humans and LLMs act in the case of ambiguity. We find that there is only a weak link between ambiguity and humans producing clarification questions in these dialogues, and low correlation between humans and LLMs. Humans hardly ever produce clarification questions for referential ambiguity, but often do so for task-based uncertainty. Conversely, LLMs produce more clarification questions for referential ambiguity, but less so for task uncertainty. We question if LLMs' ability to ask clarification questions is predicated on their recent ability to simulate reasoning, and test this with different reasoning approaches, finding that reasoning does appear to increase question frequency and relevancy.

1 Introduction

Large Language Models (LLM) are much maligned for their tendency to act presumptively, "hallucinating" in the absence of knowledge. Until the recent advent of reasoning orientated LLMs (i.e. models deliberately fine tuned with reasoning as an objective such as DeepSeek-R1 (Guo et al., 2025)), models struggled asking clarification questions and rarely proactively sought missing information (Deng et al., 2023; Li et al., 2022, 2024). Prior works have tested how LLMs respond to uncertainty, and proposed benchmarks (Zhang et al., 2024). However, this remains a challenge.

This is perhaps further complicated as clarification is a conversational strategy applied sparingly by humans (Purver et al., 2003; Rodríguez and Schlangen, 2004; Rieser and Moore, 2005). Certain situations promote greater clarification question usage; for example, situations in which information is asymmetric, and which concerns a task requiring information seeking. This happens to be a popular paradigm for tasks created with the objective of soliciting dialogue (sometimes referred to as instruction giver/instruction follower) and also, in recent years, for studying clarification questions (Chi et al., 2020; Madureira and Schlangen, 2023a; Testoni and Fernández, 2024; Shen and Lourentzou, 2023). In this work we look at the Minecraft Dialogue Corpus (Narayan-Chen et al., 2019), a task orientated, grounded corpus that follows this paradigm. We select this corpus as it has benefited from multiple separate annotation efforts (Thompson et al., 2024; Madge et al., 2025) that extend its already richly structured offering, with useful supplemental information that can inform the experiments undertaken in this work.

We focus on one particular area of linguistic uncertainty, referential ambiguity. There has been a long standing interest in reference, with ambiguity featuring as an interest in the first popular corpora (Pradhan et al., 2012). Our first contribution of this work is combining prior annotation efforts providing annotations for clarification questions (Thompson et al., 2024) and reference (Madge et al., 2025) into a single aligned corpus in the MMAX format (described in Section 3).

This annotation supports our next contribution, a comparison of how LLMs and humans resolve uncertainty. We ask, *"does referential ambiguity really trigger clarification requests from humans, and is this different for LLMs?"*. We look at both the annotated instances of linguistic ambiguity, and the original clarification questions, as posed by human interlocutors for correlation. We test both

Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2025), pages 1–11
Suzhou, China, November 9, 2025.

against different LLM based approaches.

One proposed approach to improving clarification questions with LLMs is through adding further reasoning capabilities with variations on the Chain of Thought approach (Deng et al., 2023). Despite extensive testing in epistemic, aleatoric, linguistic uncertainty (Ortega-Martín et al., 2023) and proposed benchmarks (Zhang et al., 2024), it remains somewhat unclear how effective LLMs are in identifying uncertainty and even more so how LLMs may consistently generate the relevant questions to address it.

For our second research question, we ask if the ability to ask a useful clarification question, or indeed judge when to ask a question, is based on a model's ability to simulate reasoning. We test this hypothesis with an experiment, comparing models that were trained to include reasoning and prompt engineering strategies for inducing reasoning at test time against ordinary models/methods.

For our final contribution, we look further into human reasoning and its constituent parts, with a discussion on how these may affect LLMs seemingly emergent ability to ask some clarification questions, and liken this to human reasoning.

2 Related Work

2.1 Clarification Questions

There has been a great interest in clarification questions in the literature on dialogue systems going back at least twenty years (Purver et al., 2003; Schlangen, 2004; Gabsdil, 2003). More recently, there has been extensive work in modern Natural Language Processing modelling clarification question generation or indeed when to ask them (Majumder et al., 2021; Aliannejadi et al., 2019; Kiseleva et al., 2022). This section will primarily focus on prior works that target clarification question with overlap to our specific goals (i.e. task orientated dialogue with situated and/or embodied agents). Previous works have gathered or annotated datasets in situated dialogue with clarification questions. For example, (Gervits et al., 2021) gather a corpus (HuRDL - Human-Robot Dialogue Learning) and annotate clarification questions in a dialogue gathered from human participants in a robot situated tool gathering task. (Gella et al., 2022) annotate dialogue acts in the TEACh (Task-driven Embodied Agents that Chat) dataset (Padmakumar et al., 2022); the product of a task that has human participants collaborating to perform household tasks in a virtual house environment.

A particularly popular task/dataset for this is CoDraw (Kim et al., 2017). The CoDraw task (Kim et al., 2017) is similar to the previously discussed Minecraft task, in that an instruction giver communicates with an instruction follower to collaboratively reach a goal. As opposed to constructing a 3D voxel based structure, they recreate a scene formed of clipart images. (Madureira and Schlangen, 2023b) annotate this dataset with clarification questions.

Previous works have also compared when humans and models would ask clarification questions (Testoni and Fernández, 2024) use the aforementioned CoDraw dataset to investigate this the relationship between model uncertainty, and human clarification questions based on task properties (e.g. size, orientation, position etc.). The presence of a clarification question is used as the measure of measure uncertainty, and they use logistic regression to see if they can predict this.

The Minecraft Dialogue Corpus (Narayan-Chen et al., 2019) used in this work is different, in that rather than referencing direct objects, continuously changing abstract shapes are created and manipulated during the dialogue. We expand on this further in the following section. There has been other work using Minecraft-like environments as a test-bed for the study of clarification questions in dialogue. However, this was prior to LLMs and looked at clarification question production as a task of ranking available clarification questions, rather than their generation (Kiseleva et al., 2022)

Several works have investigated the use of LLMs for clarification question generation, with methods including: fine tuning on question data (Andukuri et al., 2024); uncertainty estimation over multiple samples (Pang et al., 2024; Zhang and Choi, 2023) and multi turn prompting strategies (Kuhn et al., 2022; Li et al., 2023). To our knowledge, none of these are primarily concerned with reference or situated dialogue settings. There is however evidence to suggest LLMs can successfully resolve reference with performance similar to, or in some cases superior to, reference specific models (Hicke and Mimno, 2024; Le and Ritter, 2023).

2.2 MDC and its extensions

The Minecraft Dialogue Corpus (Narayan-Chen et al., 2019) is a collection of conversations among human participants performing the Minecraft Collaborative Building Task.This follows the typical

instruction giver, instruction follower paradigm, where *the Architect*, who has full observability over the target environment but is unable to act, instructs *the Builder*, to manipulate the environment to meet that target structure. The world is a 3D voxel based $11 \times 9 \times 11$ Minecraft like world, originally provided by project Malmo (Johnson et al., 2016). This results in a 509 multi turn situated dialogues with rich linguistic phenomena including reference and clarification.

Various annotation efforts have extended MDC, including variations of AMR (Bonn et al., 2020; Bonial et al., 2021), reference (Madge et al., 2025) and Segmented Discourse Representation Theory (SDRT, Thompson et al., 2024). We focus on the latter two as they are directly used in this work.

SDRT provides a macrostructure of interconnected logical discourse forms, linking narrative arcs and discourse relations (e.g. clarification questions, corrections, confirmations, acknowledgements etc., see Asher and Lascarides, 2003; Lascarides and Asher, 2007). Thompson et al. (2024) exhaustively annotated MDC with SDRT in their Minecraft Structured Dialogue Corpus (MSDC).

MDC-R (Madge et al., 2025) consists of a subset of 100 dialogues from MDC with reference expert annotated according to the ARRAU guidelines (Poesio et al., 2024). The dynamically changing environment and instruction based two-party dialogue gives rise to various types of reference, much of which, beyond the discourse, is linked directly to the objects in the virtual world. This results in some interesting and challenging examples of ambiguity for a dialogue system to resolve.

2.3 Reasoning in Large Language Models

Chain of Thought (COT, Wei et al., 2022) simulates reasoning at inference time by encouraging the model to think through the answer step by step. In implementation, this can take one of two common forms. The model is either provided an example of thinking through a problem step by step as part of a one-shot/few-shot prompt, or a zero-shot approach that simply prefaces the prompt with something like "Let's think this through, step by step...". The core benefits of the COT approach are given to be: problem decomposition; some explainability/insight into how results are reached; logical problem solving/symbolic manipulation and ease of application to existing models. Previous work has observed an improvement in applying this method when addressing Minecraft orientated tasks (Madge and Poesio, 2024). Following Chain of Thought, several models have been trained or aligned explicitly to follow this process (e.g. Gemini 2.5; Version 3 of the Qwen model (Yang et al., 2024); DeepSeek-R1 (Guo et al., 2025)).

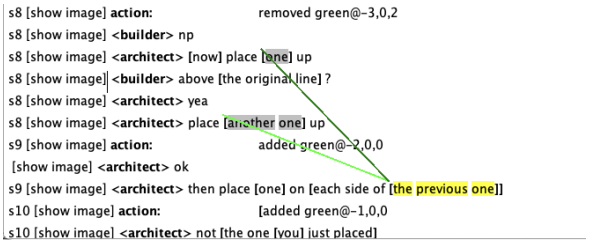

Figure 1: Referential Ambiguity Annotation Example in MMAX

3 Adding MSDC information to MDC-R

In this section we will motivate and describe our effort to add MSDC information to MDC-R to produce a new version of the corpus combining both types of annotation.

We identify two types of utterances or phrases may provoke clarification requests. Firstly, utterances that have been annotated as the subject of a clarification or confirmation request in the dialogue (typically related to task orientated uncertainty , and secondly instances of referential ambiguity.

To support our experimentation, and investigation of any possible relationship between the two, we present a corpus that merges two existing corpora that identifies these. This is a combination of the previously discussed MDC-R (Madge et al., 2025) (providing reference annotations for MDC), and MSDC corpora (Thompson et al., 2024) (providing clarification questions for MDC), permitting convenient examination of reference and more specifically types of referential ambiguity aligned with clarification questions. We automated the merge of these two corpora, through use of a script that operated at token level to produce a common method of addressing and aligning the respective segments in each.

We add a new MMAX[1] layer, referred to as SDRT. Each MMAX markable in this layer represents an Elementary Discourse Unit and the relations between those markables are represented by a *to* attribute on each markable, with the related markables unique ID. Crucially, these relationships

[1] https://mmax2.net/

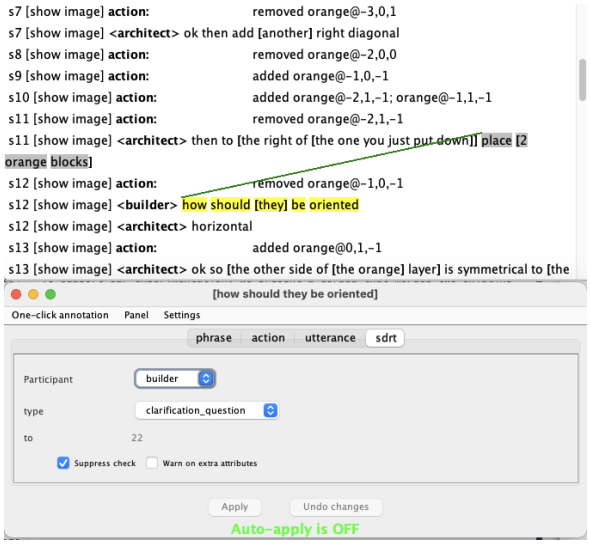

Figure 2: Clarification Question Annotation Example in MMAX

describe links between utterances and their clarification questions.

This combined format allows parsing and examination of adjacent reference in MMAX (shown in Figure 1) and clarification (shown in Figure 2).

Reference annotation has many parameters, with each relationship holding many attributes. Whilst the SDRT annotations exhaustively cover all of the original MDC dataset (Narayan-Chen et al., 2019), MDC-R covers a subset of 100 dialogues (some more detailed descriptive statistics taken from MDC-R (Madge et al., 2025) are given in Table 1). As such, our corpus will be limited to the same 100 dialogues.

Statistic	Count	Statistic	Count
Documents	101	Tokens	29,174
Utterances	3,343	Actions	5,793
Markables	7,600	Discourse old	1960
Bridging	1,053	Object	500
Plural	24	Ambiguous	149

Table 1: MDC-R Corpus Statistics (Madge et al., 2025)

Using the combined corpus, we have counted instances of utterances exhibiting certain attributes that may motivate a question. Firstly, using annotations originating from MSDC, we look at the counts of *confirmation questions* and *clarification questions* occurring in the original human dialogue. Secondly, we count specific instances of referential ambiguity, using annotations originating from MDC-R. These are instances of discourse deixis - relating to parts of the discourse (e.g. "as I said earlier"), and spatial deixis in a real world space - typically our voxel world environment (e.g. "next to that block").

Table 2 shows the frequency of these different instances of utterances that may motivate a question as a percentage of all utterances in the selected subset of the corpus.

Type	Instances	%
confirmation question	218	3.7%
clarification question	182	3.1%
discourse deixis	24	0.4%
spatial deixis	16	0.3%

Table 2: Frequency of instances as a percentage of utterances

77% of dialogues contain a clarification question and 75% a confirmation question.

The SDRT annotations of the complete corpus found in MSDC had 999 confirmation questions and 960 clarification questions over 547 dialogues. We can see from the relative quantity of questions, that the selected dialogues do appear to be representative of the corpus as a whole, with respect to question quantity.

To give some overview, the most common phrases for discourse deixis are: *"that"* (8); *"this"* (7) and *"it"* (3). Spatial deixis has 3 instances of "this", but referents while still ambiguous, tend do be more literal (e.g. *"the red end"*). We expand further on this in a discussion of reoccurring patterns in Section 5.1.

The corpus is available at `https://github.com/arciduca-project/MDC-R/tree/sdrt`.

4 Methodology

To test how LLMs perform clarification questions, we first discover points in the conversation that may require clarification, then we sample from various LLMs with different approaches, feeding the context of the conversation up to the appropriate point. Examples of our prompts are given in the Appendix (see Section A.1). A system prompt describes the nature of environment and it's constraints. This differs slightly between architects and builders, in that architects can see the target structure, and the builder's system prompt specifies the required JSON response format necessary to encode their resultant actions or expect the world state (this experiment is text

only, with the world state/actions encoded in JSON - no images are used). For the chain-of-thought treatments, we supplement these system prompts as follows. The zero-shot chain-of-thought approach simply adds, "Think step by step" to the system prompt. The one-shot approach follows the system prompt with an example exchange that incorporates thinking.These approaches are deliberately selected to compare reasoning based models (e.g. `llama3.2`, `DeepSeek-R1` (Guo et al., 2025)) and sampling methods (`llama2:13b-COTZERO`, `llama2:13b-COTONE`) vs. non-reasoning (`llama2:13b`). We compare the approach taken by LLMs to the approach originally taken by humans.

There are three characteristics of the instances that we identify to test against. Firstly, the subject of any clarification and confirmation question as originally annotated in SDRT. Second and thirdly, linguistic ambiguities. We select any referent that has two or more antecedents, whether they be discourse deixis (part of the discourse), or spatial deixis (in reference to objects in the environment. That is to say, in the MDC-R MMAX format, for discourse deixis the phrase in question, would have the following attributes specified `segment_phrase_antecedent_2`, and for spatial deixis, `object2` specified.

We also check for correlation between clarification questions, and the instances of linguistic ambiguity, as permitted by using our new merged corpus (described in Section 3),

We measure the tendency of different approaches to ask a question and use a single human coder to examine specific instances of ambiguity and their responses to attempt to quantify the number of relevant questions asked when applying each method. Prior to quantifying responses to instances we conduct a thorough example driven investigation.

5 Results

In the first section we look at how LLMs and humans respond to instances of referential linguistic ambiguity, including whether humans pose clarification questions when they encounter linguistic ambiguity. In the second section we look at the original clarification questions as posed by humans, to see how LLMs respond. Finally, we perform a quantitative evaluation, counting the tendency of different approaches to ask questions, and a count of question relevancy under one specific condition.

5.1 Linguistic Ambiguities

Do humans ask clarification questions when there is ambiguity? In many cases, where linguistic ambiguity exists, it appears deliberately underspecified in the interest of brevity, and where expanding more literally or verbosely would not have any further positive effect on task completion. Here, we see two common patterns emerge.

The first common pattern is in the communication of approval. The exact target of approval is unspecified, and arguably redundant if correct, as no further intervention is required. This appears to be used as a signal to indicate objective completion and generally precede new instruction. Examples include: *"oh that is magnificent builder"*; *"... that is great"*; *"...that's perfect"*; *"...let 's see what we can do about that"*; *"yeah, that's fine"*; *"good job"*, *"yes like that"*.

These seem closely related to the second commonly appearing pattern in sources of linguistic ambiguity, which occur in clarification questions themselves and appear to be used as a polite conversational device to invite architect interruption to general correction and indicate the builder believes the prior instruction to be addressed. Examples include: "*like this?*"; "*is this good?*"; "*this right?*", where "this" and "that" do not have clear referrents. Again, exact specification of what "this" refers to is somewhat redundant, as any corrective instruction may follow. Examining the context an interpreting more explicitly, these would appear to be more semantically equivalent to: "I believe I've completed all instructions, I'm ready for the next instruction or correction".

The first finding of this paper is that linguistic ambiguities are rarely the target of clarification questions directly in our corpus. We only discovered a single instance in which a discourse deixis referential ambiguity is in the vicinity of the subject of a latter clarification question. In the case of linguistic ambiguity, it seems humans prefer to act presumptively, in this setting.

Do LLMs ask clarification questions when there is ambiguity? Unresolved linguistic ambiguities, left unchecked, do propagate, resulting in the necessity for clarification or confirmation later in the conversation. Whilst human participants do not address these immediately, LLMs sometimes do. We show two examples here. In the following example, "the previous one" has an ambiguous referent in the context of the conversation and current world state.

This is not immediately resolved at the time by the human builder, who acts presumptively, and takes action. Having taken the incorrect action, and then been required to perform a revision, the builder then uses a confirmation question, *"like that?"*. In contrast, the LLM however, does ask a question to directly disambiguate, *"the previous one"* by offering the two possible interpretations (see Figure 3).

Architect	then place one on each side of the previous one
Builder (model vs. original human dialogue):	
Model	Are the blocks to be placed on each side of the green block at [-2,0,0] or in relation to the vertical stacking process near the center?
Human	*takes action*
Architect	not the one you just placed
Builder	*revisits with corrective action*
Builder	like that?

Figure 3: Excerpt from B4-A19-C4-1522882189483 demonstrating propagation of referential ambiguity

Similarly, in Figure 4, the human builder does not clarify the ambiguous statement, "the same thing", but proceeds to take action, to discover a lot of work must then be mirrored. Conversely, an LLM asks, about which plane the blocks should be mirrored, to first discover the correct approach.

Builder	*builder takes action*
Architect	now the same thing on the other side
Builder (model vs. original human dialogue):	
Model	Does 'the same thing on the other side' mean mirroring the blocks symmetrically with respect to the y-z plane at x=0?
Human	*builder takes further action*
Architect	that but on the other side
Builder	*builder carries out extensive revision*
Architect	there

Figure 4: Excerpt from B35-A34-C118-1524004322117 demonstrating direct and immediate clarification of linguistic ambiguity from LLM

We show here in this section that whilst linguistic ambiguities rarely trigger direct human clarification, despite the fact they may translate to uneconomical task action, LLMs can and do sometimes respond with targeted clarification questions.

5.2 Clarification/Confirmation Questions

Having shown in the previous section, there is not a strong correlation between linguistic ambiguity and clarification questions, this section looks, at under what circumstances clarification questions are asked by human participants, and how this behaviour compares with how LLMs respond.

A common repeated pattern by human builders is the use of questions such as, *"like that?"*, as previously discussed, and others that are potentially more specific, such *"is this right?"* and *"how's that"*, with contextually evident reference to specific actions. This is not a behaviour employed by the LLMs tested here. As such, these clarification questions are unique to human participants.

Another clarification question approach exhibited uniquely by humans, is that of interrupting a set of instructions to proactively anticipate a possible pattern that they have observed to potentially reduce further instruction. For example: *"Are we repeating the structure?"*; *"Are we repeating the pattern with orange and yellow?"*; *"both sides or just this one?"*; *"are we going to repeat what we did on the other side?"*; *"same design? just blue?"*; *"repeat here?"*.

Clarification questions typically occur not directly around linguistic ambiguity, but rather ambiguity in relation to the task objectives (e.g. properties). In the following example, the architect specifies colour and quantity, but does not specify orientation, leading to similar, albeit more concise on part of the human, questions from human and LLM (Figure 5). Similarly, for individual block position in Figure 6.

Architect	then to the right of the one you just put down place 2 orange blocks
Builder (model vs. original human dialogue):	
Model	What is the current orientation of the blocks that have been placed so far? Are they forming a diagonal line from north to south or from east to west?
Human	how should they be oriented

Figure 5: Excerpt from B19-A36-C118-1524093668453 showing the subject of a clarification request in relation to orientation uncertainty

Architect	great now one orange below that one
Builder (model vs. original human dialogue):	
Model	Is it correct to assume 'now one orange below that one' means in the same diagonal direction as 'now one red block down along the diagonal
Human	diagonally or nah?

Figure 6: Excerpt from B16-A29-C144-1524067263550 showing the subject of a clarification request in relation to individual position uncertainty

There are several examples in relation to collective, individual, absolute and relative property. LLMs seem to also approach these well. Another common example is colour (shown in Figure 7).

Again, the LLM reaches a fair clarification question.

Architect	on the yellow block touching the orange , add two blocks to either side of it . making a t.
Builder (model vs. original human dialogue):	
Model	What color are the new blocks to form the T shape and do you want me to remove or replace any existing blocks at [-4,1,0], [-3,0,0], [-2,1,0]?
Human	what color blocks ?

Figure 7: Excerpt from B37-A23-C102-1523629957597 showing the subject of a clarification request in relation to the colour property

5.3 Quantitative Analysis

In this section we aim to quantify the differences between approaches and how they respond to situations that may benefit from clarification. Primarily, we are interested in each individual approaches tendency to generate a question (Table 3), particularly a relevant questions. (Table 4)

Table 3 looks at the number of questions asked in each reply. In this first case, we do not assess question relevance, but simply the tendency to reply with a question. We see a lower tendency to ask a question in the previous generation models that did not feature reasoning (i.e. `llama2`). However, later models that do feature reasoning (e.g. `llama3` and `deepseek-r1`), or previous generation models supplemented with chain-of-thought prompts (e.g. `llama2:13b-COTZERO` and `llama2:13b-COTONE`) that attempt to simulate reasoning, perform comparably with a greater tendency to question.

Approach	Questions Asked
llama2:13b	276
llama2:13b-COTZERO	303
llama2:13b-COTONE	327
deepseek-r1:8b	354
llama3.2:3b	278
llama3.3:70b	383

Table 3: Number of questions asked in different approaches

To provide some notion of relevancy, Table 4 looks specifically at the number of questions asked that target the instance of expert annotated ambiguity compared with the actions of the original human participant in the conversation. These are counted solely for instances of spatial deixis, as these are very literal and therefore the easiest to objectively assess. For example, in *"now a tower of five oranges on top of the red end"*, Llama3.3:70b's response of *"Which end of the red blocks is considered the 'red end', the one at coordinates [3, 0, 1] or [0, 0, 1]?"*, is considered to target the ambiguity.

We see that the previous generation models, not trained for reasoning, but with Chain of Thought prompting, perform comparably to modern reasoning orientated models, in this regard. Human participants did not choose to disambiguate these phrases direct at the time, but rather acted presumptively. This approach does not measure the final utility of asking the question, or any impact it may have on the conversation.

Approach	Spatial Deixis
Human participant	0
llama2:13b	0
llama2:13b-COTZERO	1
llama2:13b-COTONE	3
deepseek-r1:8b	3
llama3.2:3b	2
llama3.3:70b	5

Table 4: Number of relevant spatial deixis questions asked by approach

Another limitation to our experiment. It's challenging to communicate a sense of perspective to the LLMs. As a consequence, some instructions do not make sense, e.g.: *"A: can you come to the side of the structure so you have a side view"*; *"B: left or right"*; *"B: forward to my right or in front of me?"*; *"B: this perspective ?"*.

6 Reasoning and Clarification

Our experiments would appear to show that the recent advent of reasoning in models has the emergent benefit of allowing models to ask clarification questions. In this section, following our hypothesis that clarification is dependent on an ability to reason, we look at clarification questions in relation to reasoning through the lens of human psychology, and where available, assessments of LLM abilities to perform the required components of related human reasoning abilities

Examining the role of reasoning in clarification, knowing when to ask a question, requires reflection on the gaps in ones own knowledge, or a higher order of thought, referred to in human psychology as **Metacognition** (Flavell, 1979). LLMs originally lacked any awareness of gaps in their knowl-

edge, acting presumptively, leading to "hallucination" where, albeit often grammatically valid, a model's output would be factually incorrect or possibly nonsensical. This is most commonly due to a lack of knowledge (Zhang et al., 2023), or perhaps the inability to reason when knowledge is absent. Metacognition is the ability to reflect on held knowledge by self questioning. The Metacognitive capabilities of LLMs have been explored in previous works in relation to reasoning (Didolkar et al., 2024). Metacognitive prompting that, self questions to enhance reasoning (Wang and Zhao, 2023) has been explored in LLMs, as has self-questioning with the goal of reducing hallucination (Dhuliawala et al., 2023). A model knowing whether it has the applicable knowledge or skill to proceed, or whether to direct its process to clarification, could be seen as metacognitive regulation.

When an agent is working in collaboration with other interlocutors, clarification may be dependent on discovering other parties knowledge, abilities or attitudes and approach to a task. This crucial component, in psychology, is an aspect of reasoning known as **Theory of Mind** (Premack and Woodruff, 1978), which relates to reasoning about other participants belief states. There is some evidence to suggest, as a consequence of simulated reasoning ability, LLMs may now be able to simulate this also (Kosinski, 2023). This has been explored with LLMs in Mindcraft, which is a collaborative task in which the players have separate skills and must negotiate to reach a common goal (Bara et al., 2021). This is particularly important for LLMs in referential communication (Sidera et al., 2018)

7 Conclusion

To conclude, we find that in our conversations human participants do not commonly ask a clarification question when language is ambiguous. In the majority of these cases of linguistic ambiguity, the ambiguity appear deliberate in the interest of conversational efficiency. Consequently, there may be little utility to asking a question in many of those situations. (The resultant utility of a clarification question is not examined in this work, and may be the subject of a future work.) We did however identify one situation in which that uncertainty did propagate to create future issues. Regardless of conversational efficiency, we do find that LLMs, particularly reasoning orientated ones, are capable of asking relevant clarification questions under those circumstances.

On the topic of the originally posed human clarification questions, we found they largely followed a specific pattern and strategy not adopted by the LLMs. That was, to perform actions then use a clarification question to verify they were correct. These instances of human clarification questions largely relate to task based ambiguity rather than linguistic ambiguity.

Across all instances, we find a greater tendency of reasoning orientated approaches to pose clarification questions and find that this can be somewhat matched at test time, with methods such as COT.

Acknowledgements

This research was funded by ARCIDUCA, EPSRC EP/W001632/1

References

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 475–484.

Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D Goodman. 2024. Star-gate: Teaching language models to ask clarifying questions. *arXiv preprint arXiv:2403.19154*.

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. 2021. Mindcraft: Theory of mind modeling for situated dialogue in collaborative tasks. *arXiv preprint arXiv:2109.06275*.

Claire Bonial, Mitchell Abrams, David Traum, and Clare Voss. 2021. Builder, we have done it: evaluating & extending dialogue-amr nlu pipeline for two collaborative domains. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 173–183.

Julia Bonn, Martha Palmer, Jon Cai, and Kristin Wright-Bettner. 2020. Spatial AMR: Expanded spatial annotation in the context of a grounded minecraft corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020),*.

Ta-Chung Chi, Minmin Shen, Mihail Eric, Seokhwan Kim, and Dilek Hakkani-Tur. 2020. Just ask: An interactive learning framework for vision and language navigation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 2459–2466.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. *arXiv preprint arXiv:2305.13626*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *Preprint*, arXiv:2309.11495.

Aniket Didolkar, Anirudh Goyal, Nan Rosemary Ke, Siyuan Guo, Michal Valko, Timothy Lillicrap, Danilo Jimenez Rezende, Yoshua Bengio, Michael C Mozer, and Sanjeev Arora. 2024. Metacognitive capabilities of llms: An exploration in mathematical problem solving. *Advances in Neural Information Processing Systems*, 37:19783–19812.

John H Flavell. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10):906.

Malte Gabsdil. 2003. Clarification in spoken dialogue systems. In *Proceedings of the 2003 AAAI Spring Symposium. Workshop on Natural Language Generation in Spoken and Written Dialogue*, pages 28–35.

Spandana Gella, Aishwarya Padmakumar, Patrick Lange, and Dilek Hakkani-Tur. 2022. Dialog acts for task-driven embodied agents. *arXiv preprint arXiv:2209.12953*.

Felix Gervits, Antonio Roque, Gordon Briggs, Matthias Scheutz, and Matthew Marge. 2021. How should agents ask questions for situated learning? an annotated dialogue corpus. *arXiv preprint arXiv:2106.06504*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Rebecca MM Hicke and David Mimno. 2024. [lions: 1] and [tigers: 2] and [bears: 3], oh my! literary coreference annotation with llms. *arXiv preprint arXiv:2401.17922*.

Matthew Johnson, Katja Hofmann, Tim Hutton, and David Bignell. 2016. The malmo platform for artificial intelligence experimentation. In *IJCAI*, volume 16, pages 4246–4247.

Jin-Hwa Kim, Nikita Kitaev, Xinlei Chen, Marcus Rohrbach, Byoung-Tak Zhang, Yuandong Tian, Dhruv Batra, and Devi Parikh. 2017. Codraw: Collaborative drawing as a testbed for grounded goal-driven communication. *arXiv preprint arXiv:1712.05558*.

Julia Kiseleva, Ziming Li, Mohammad Aliannejadi, Shrestha Mohanty, Maartje ter Hoeve, Mikhail Burtsev, Alexey Skrynnik, Artem Zholus, Aleksandr Panov, Kavya Srinet, and 1 others. 2022. Interactive grounded language understanding in a collaborative environment: IGLU 2021. In *NeurIPS 2021 Competitions and Demonstrations Track*, pages 146–161. PMLR.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 4:169.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2022. Clam: Selective clarification for ambiguous questions with generative language models. *arXiv preprint arXiv:2212.07769*.

Alex Lascarides and Nicholas Asher. 2007. Segmented discourse representation theory: Dynamic semantics with discourse structure. In *Computing meaning*, pages 87–124. Springer.

Nghia T. Le and Alan Ritter. 2023. Are large language models robust coreference resolvers? *Preprint*, arXiv:2305.14489.

Belinda Z Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. 2023. Eliciting human preferences with language models. *arXiv preprint arXiv:2310.11589*.

Haau-Sing Li, Mohsen Mesgar, André FT Martins, and Iryna Gurevych. 2022. Asking clarification questions for code generation in general-purpose programming language. *arXiv preprint arXiv:2212.09885*.

Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei W Koh, and Yulia Tsvetkov. 2024. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888.

Chris Madge, Maris Camilleri, Paloma Carretero Garcia, Mladen Karan, Juexi Shao, Prashant Jayannavar, Julian Hough, Benjamin Roth, and Massimo Poesio. 2025. Mdc-r: The minecraft dialogue corpus with reference. *arXiv preprint arXiv:2506.22062*.

Chris Madge and Massimo Poesio. 2024. A llm benchmark based on the minecraft builder dialog agent task. In *Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue*.

Brielen Madureira and David Schlangen. 2023a. " are you telling me to put glasses on the dog?"content-grounded annotation of instruction clarification requests in the codraw dataset. *arXiv preprint arXiv:2306.02377*.

Brielen Madureira and David Schlangen. 2023b. Instruction clarification requests in multimodal collaborative dialogue games: Tasks, and an analysis of the codraw dataset. *arXiv preprint arXiv:2302.14406*.

Bodhisattwa Prasad Majumder, Sudha Rao, Michel Galley, and Julian McAuley. 2021. Ask what's missing and what's useful: Improving clarification question generation using global knowledge. *arXiv preprint arXiv:2104.06828*.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5405–5415.

Miguel Ortega-Martín, Óscar García-Sierra, Alfonso Ardoiz, Jorge Álvarez, Juan Carlos Armenteros, and Adrián Alonso. 2023. Linguistic ambiguity analysis in chatgpt. *arXiv preprint arXiv:2302.06426*.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. 2022. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2017–2025.

Jing-Cheng Pang, Heng-Bo Fan, Pengyuan Wang, Jia-Hao Xiao, Nan Tang, Si-Hang Yang, Chengxing Jia, Sheng-Jun Huang, and Yang Yu. 2024. Empowering language models with active inquiry for deeper understanding. *arXiv preprint arXiv:2402.03719*.

Massimo Poesio, Maris Camilleri, Paloma Carretero Garcia, and Ron Artstein. 2024. *The ARRAU 3 Annotation Manual*, v. 1.1 edition. Queen Mary University of London.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint conference on EMNLP and CoNLL-shared task*, pages 1–40.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. 2003. On the means for clarification in dialogue. *Current and new directions in discourse and dialogue*, pages 235–255.

Verena Rieser and Johanna Moore. 2005. Implications for generating clarification requests in task-oriented dialogues. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 239–246, Ann Arbor.

Kepa Rodríguez and David Schlangen. 2004. Form, intonation and function of clarification requests in German task-oriented spoken dialogues. In *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue (Catalog)*, pages 101–108, Barcelona.

David Schlangen. 2004. Causes and strategies for requesting clarification in dialogue. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 136–143.

Ying Shen and Ismini Lourentzou. 2023. Learning by asking for embodied visual navigation and task completion. *arXiv preprint arXiv:2302.04865*.

Francesc Sidera, Georgina Perpiñà, Jèssica Serrano, and Carles Rostan. 2018. Why is theory of mind important for referential communication? *Current Psychology*, 37:82–97.

Alberto Testoni and Raquel Fernández. 2024. Asking the right question at the right time: Human and model uncertainty guidance to ask clarification questions. *arXiv preprint arXiv:2402.06509*.

Kate Thompson, Julie Hunter, and Nicholas Asher. 2024. Discourse structure for the minecraft corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4957–4967.

Yuqing Wang and Yun Zhao. 2023. Metacognitive prompting improves understanding in large language models. *arXiv preprint arXiv:2308.05342*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Michael JQ Zhang and Eunsol Choi. 2023. Clarify when necessary: Resolving ambiguity through interaction with lms. *arXiv preprint arXiv:2311.09469*.

Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. 2024. Clamber: A benchmark of identifying and clarifying ambiguous information needs in large language models. *arXiv preprint arXiv:2405.12063*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

A Appendix

A.1 System Prompts

Architect You are an agent in a voxel world, where the most northernly point is 0,0,-5; the most westerly point -5,0,0; the most eastern point is 5,0,0; the most southern 0,0,5 and the y-axis is

up and down, with y=0 being the minimum. Your task is deliver instructions and respond to questions where necessary, to achieve a target world state []. The current world state is []. They should be easy to understand instructions, Do not ask for specific coordinates and do not accept them. Omit any pleasantries. You may ask questions

Builder You are an agent in a voxel world, using the following coordinate system, where the center on the ground is [0,0,0], [0,9,0] is the highest point in the air, the most northernly point is [0,0,-5]; the most westerly point [-5,0,0]; the most eastern point is [5,0,0]; the most southern [0,0,5]. Describe the coordinates of the blocks their colours (must be one of: blue, yellow, green, orange, purple, red) and whether the action is to add or remove them. You MUST include a question if the instruction is potentially unclear or ambiguous and you MUST follow the final JSON format: "add": [[x,y,z,"color"], ...], "remove": [[x,y,z,"color"], ...], "confidence": 0.0, "question": "...". The current world state is []. Give your final response in JSON only, without additional dialog. You may ask questions

A.2 COTONE

user

- builder: hi, let's start
- architect: place three blue blocks in a row in the center. Place a red block on top of the blue block

assistant

- architect: **Thinking:** let's think. we know three blue blocks need to go in the center. The center is [0,0,0], so to place three blue blocks would be [-1,0,0,"blue"],[0,0,0,"blue"],[1,0,0,"blue"]. Now we need to place a red block on top of the blue block. Wait, they didn't tell us which blue block, there are three. We must ask a question also. "Which of the three blue blocks would you like the red block placed on?". We need to formulate this response in the given JSON format.
- **Response:** builder: {"add": [[-1, 0, 0, "blue"], [0, 0, 0, "blue"], [1, 0, 0, "blue"]], "confidence": 0.5, "question": "Which of the three blue blocks would you like the red block placed on?"}