Cognitive Aspects of Avatars
LLM-Powered Assistants in XR/VR environments

Oier Lopez de Lacalle
University of the Basque Country (EHU)

Work done in collaboration

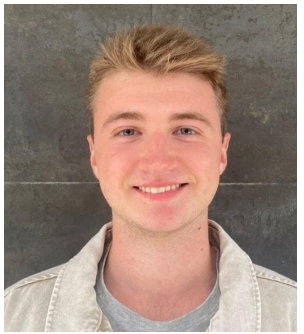

Jeremy Barnes


Ander Salaberria


Gorka Azkune


Iñigo Alonso


Markel Ferro


Oier Ijurko

EHU | HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

Main goal of the talk

To discuss the use of LLMs in virtual environments and to demonstrate (empirically) their limitations and the importance of grounding.

Outline of the talk

- LLMs as Chatbots/Assistants
- Multimodal Contexts and Grounding
  - BIM-LLM: Coding based Grounding using In-Context-Learning [EMNLP'25, submitted]
  - Agentic Approach: ReAct as a Task-Based Dialog System
- Limitation of Current Multimodal LLMs
  - The MATE benchmark [ACL'25]
  - Reasoning for Coreference Resolution in Task-Based Dialogue Systems

LLMs as Chatbots

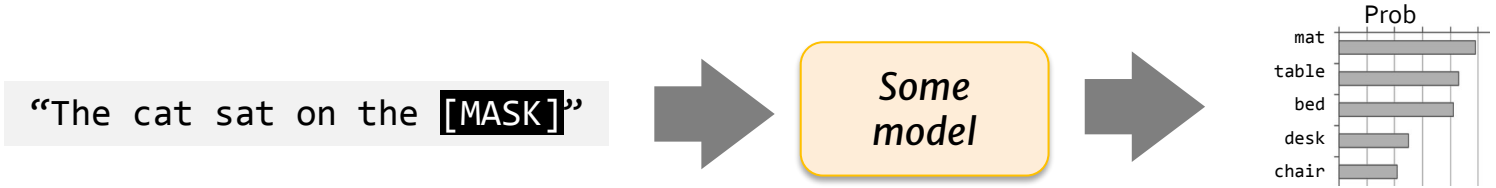
What are LLMs?

- Trained on massive text data predicting the next word given the context.
- Understands and generates human-like language.
- Based on transformer architectures.
- Powers tools like chatbots, summarizers, assistants.

next word

context

$$\mathbf{P}(X_t \mid X_1, \ldots, X_{t-1})$$



Pre-training, Instruction, and Alignment in LLMs

1. Pre-training 🏗️

- Train the model on massive amounts of text (books, articles, code, web).
- Learns general patterns of language, grammar, and world knowledge.
- Objective: predict the next word (self-supervised learning).
- **Outcome**: broad but raw capability (not yet task-specific).

2. Instruction Tuning 🎯

- Fine-tune the model on datasets with instructions and examples.
- Teaches the model to follow user prompts more directly.
- Examples: “Translate this text,” “Summarize this paragraph.”
- **Outcome**: better responsiveness to explicit instructions.

3. Alignment 🤝

- Human evaluators rank responses for helpfulness, safety, and correctness.
- Model is adjusted to prefer aligned behaviors.
- Adds guardrails against harmful, biased, or irrelevant outputs.
- **Outcome**: safer, more useful, and human-preference–aligned AI.

Chatbot = LLM + Knowledge + Memory

Knowledge Integration

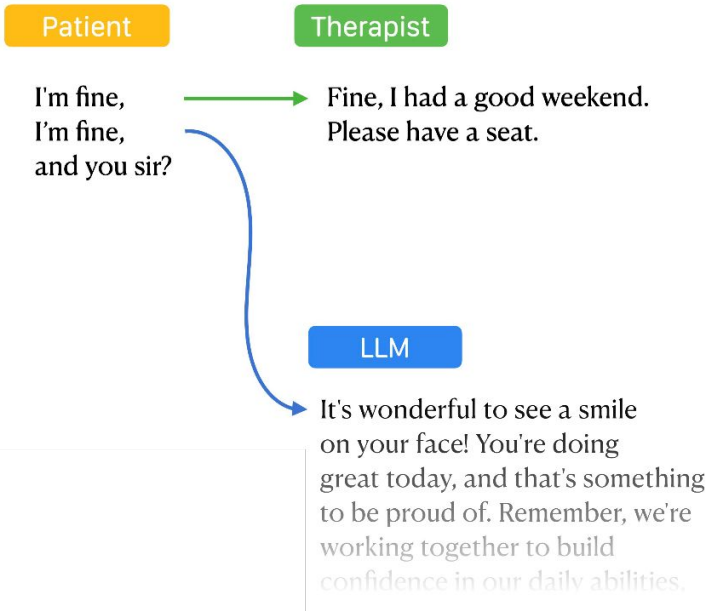
- Connect to domain-specific data (via retrieval-augmented generation, databases, or embeddings).
- Keep responses up to date with external sources when required.

Conversation Design

- Build dialogue flows (memory, context handling, multi-turn reasoning).
- Control tone and style to match your brand or use case.

Chatbots Unaware of the Environment

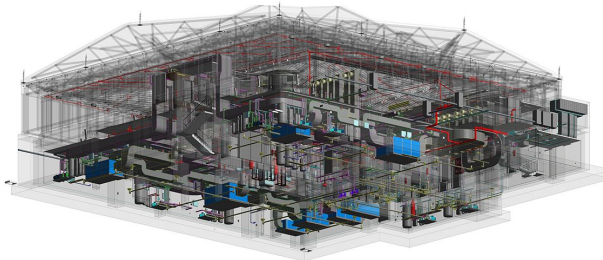
- **Objective**: responses need to be aware of the environment
- **Limitation** of agnostic approaches (e.g. neurorehabilitation):



Multimodal Contexts and Grounding

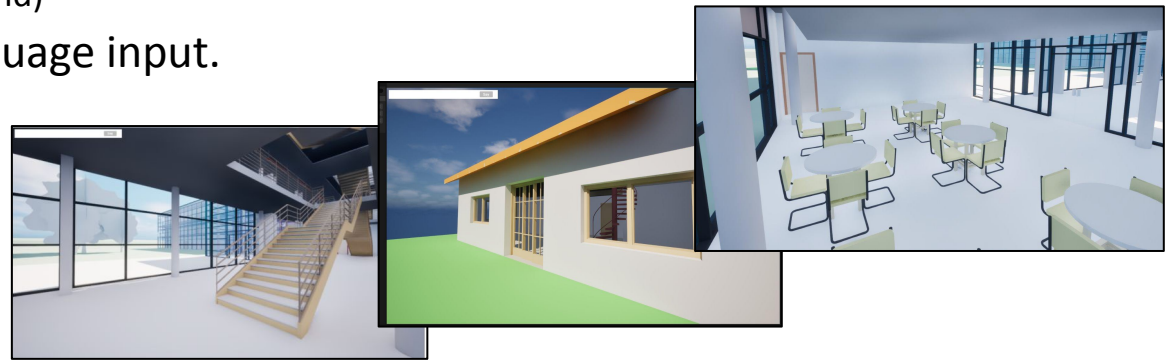
BIM-LLM: A Virtual Assistant for Architectural Design in a VR Environment [EMNLP25, submitted]

- Architectural design relies on 3D modeling procedures using **Building Information Modeling** (BIM) formats.
- BIM environments demand **specialized knowledge**.
- Changes need to be **implemented manually**, lengthening the design process and making it difficult quick prototyping.
- Incorporating an **LLM assistant** that is able to **answer queries** and **make changes directly in the VR scenario** would allow for quicker prototyping and streamline the design process.



BIM-LLM: The Virtual Assistant

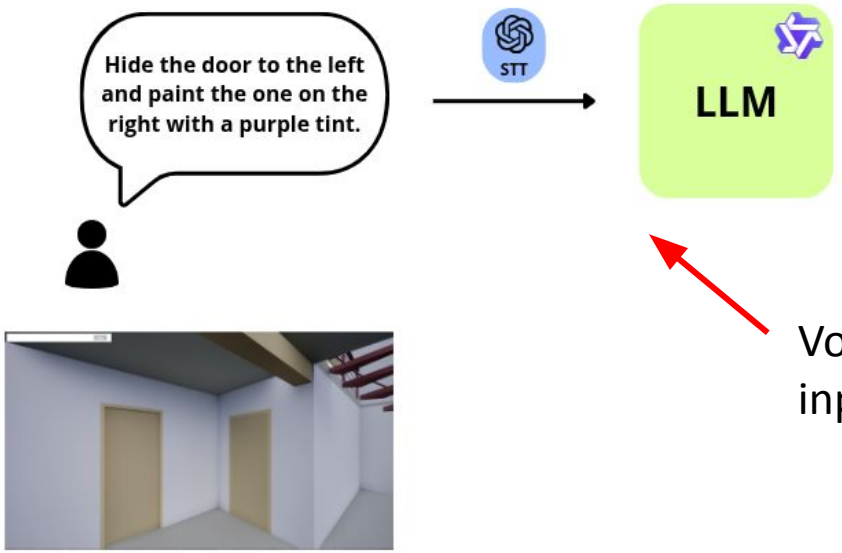
- BIM-LLM allows a BIM user to make changes directly in a VR environment via voice commands (“*paint all the window frames in blue*”) .
- Enables an LLM to:
  - Reason over spatial relationships
  - Perform multi-step operations (object manipulation, changes to object visibility)
  - Camera control (moving around)
- Based on user’s natural language input.



Demo:
https://www.youtube.com/watch?v=80whykBeR0w

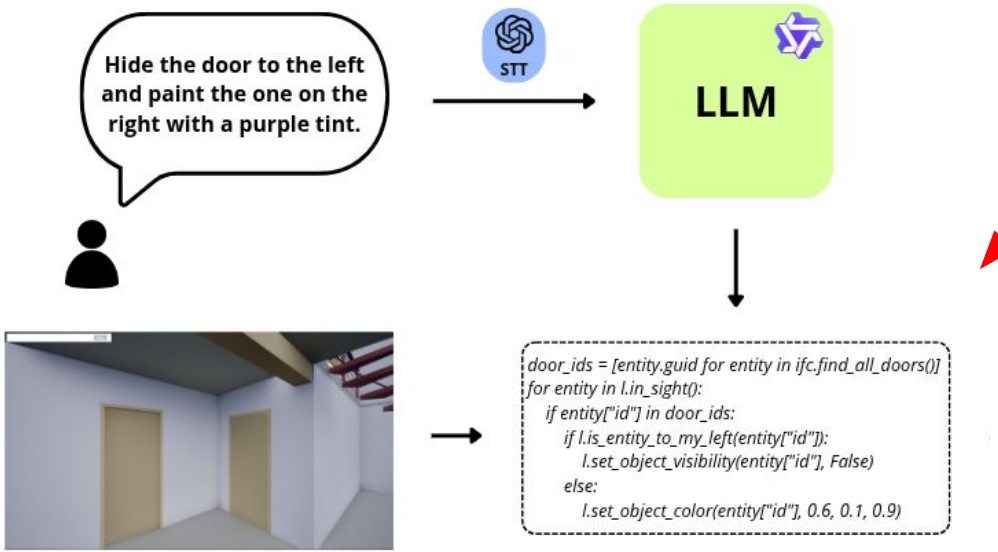
BIM-LLM: System Architecture

BIM-LLM: System Architecture



Voice is transcribed to text as input for LLM that generates code

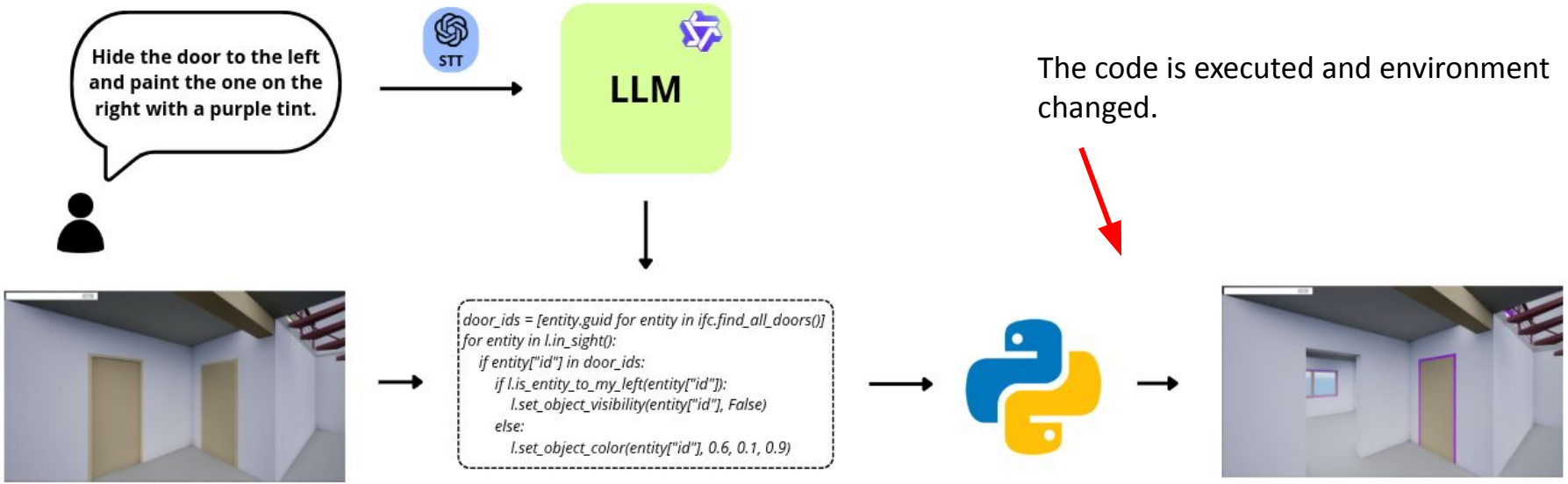
BIM-LLM: System Architecture



LLM as an interpreter of the user's intent and translates their query into executable Python code:

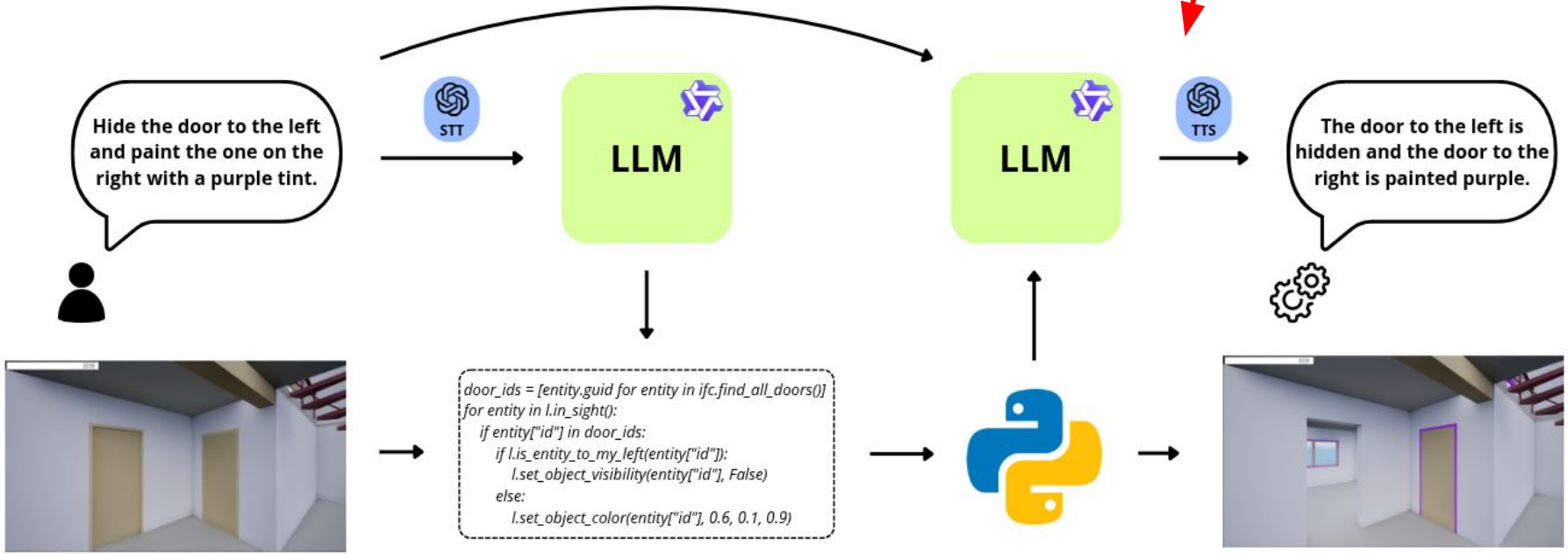
1. Identifies all doors in the building.
2. Checks which doors are in sight.
3. Hides the one that is to the user's left and paints the one to the right in purple.

BIM-LLM: System Architecture



BIM-LLM: System Architecture

The LLM takes the signal that the code has been executed without runtime errors and provides feedback to the user.



BIM-LLM: Prompt for Code Generator

The prompt incorporates richer contextual information for correct coding:

- **Task definition**: Specifies what information is available and the output format.
- **API documentation**: Defines the available Python functions and classes.
- **Initial State of the BIM**: The initial configuration of the environment.
- **Few-shot examples**: A set of seven query-code pair examples.
- **User query**: The specific input provided by the user.

BIM-LLM: Evaluation

- **Dataset**: 60 instances, including visibility, coloring, transformation, removal of BIM objects.

Model	Error Type ↓		Accuracy ↑
	Runtime	Semantic	
Qwen2.5-1.5B	40.0	**43.3**	17.7
Qwen2.5-7B	26.7	49.6	23.7
Qwen2.5-32B	20.0	44.3	35.7
Qwen2.5-72B	**11.7**	50.6	**37.7**

Table 1: Percentage of instances that were correctly completed. Incorrect instances are separated between the ones that failed to finish the execution (runtime) and the ones with incorrect outcomes (semantic).

BIM-LLM: Evaluation

- **Dataset**: 60 instances, including visibility, coloring, transformation, removal of BIM objects.

Model	Error Type ↓		Accuracy ↑
	Runtime	Semantic	
Qwen2.5-1.5B	40.0	**43.3**	17.7
Qwen2.5-7B	26.7	49.6	23.7
Qwen2.5-32B	20.0	44.3	35.7
Qwen2.5-72B	**11.7**	50.6	**37.7**

Table 1: Percentage of instances that were correctly completed. Incorrect instances are separated between the ones that failed to finish the execution (runtime) and the ones with incorrect outcomes (semantic).

- Accuracy improves with model size
- Smaller models struggle following API specifications
- Still accuracy is quite low in largest model (72B)
- Margin of improvement:
  - Better grounding is key for improving effectiveness

Agentic Approach: **ReAct** as a Task-Based Dialog System

- LLM often generate false information, resulting in unreliable results if deployed.

User: *Hello. Are there any coats by Cats Are Great?*

A simple not aligned LLM

System: *There are many coats in the world and some of them are made by Cats Are Great. Coats are usually found in...*

Aligned LLM

System: *Tell me if you like the orange coat on the table. It's by Cats Are Great.*

Correct form, but incorrect information

Aligned & Factual

System: *Tell me if you like the black coat on the back right wall, top row. It's by Cats Are Great.*

We need to ground the model

Avoid hallucinations in responses

[Yao et al. ReAct: Synergizing Reasoning and Acting in Language Models ICLR, 2023]

Prompting : “Though⟶Act ⟶ Observe”
Act = Tool use for accessing metadata

+

What’s the price for chair?

I need to know which chair

Look[]

{‘type’: ‘chair’, ‘model’: 1242, …}

Knowing the model, I want to search for the price

Search[model = 1242]

{‘model’: 1242, ‘price’: ‘**$125**’, …}

Now that I have the price I can answer

The price for the chair is **$125**.

User

LLM - reasoning

Observation - external APIs

LLM - Acting

LLM - Output

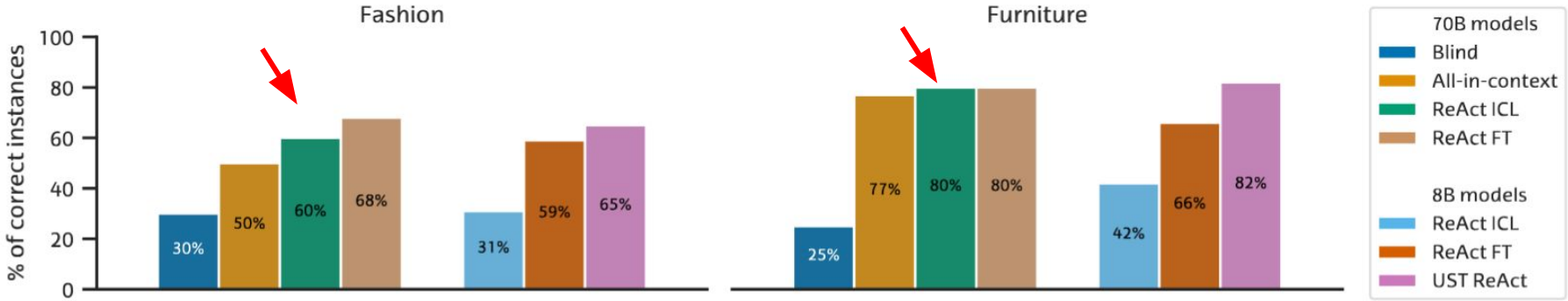
LUMINOUS
LANGUAGE AUGMENTATION FOR HUMANVERSE

Universidad del País Vasco | Euskal Herriko Unibertsitatea | DE INFORMATICA

Applying ReAct to SIMMC2.1

- Develop an agentic chatbot that acts as a salesperson.
- **Manual Evaluation**: Answers must be factually correct and relevant to the question.

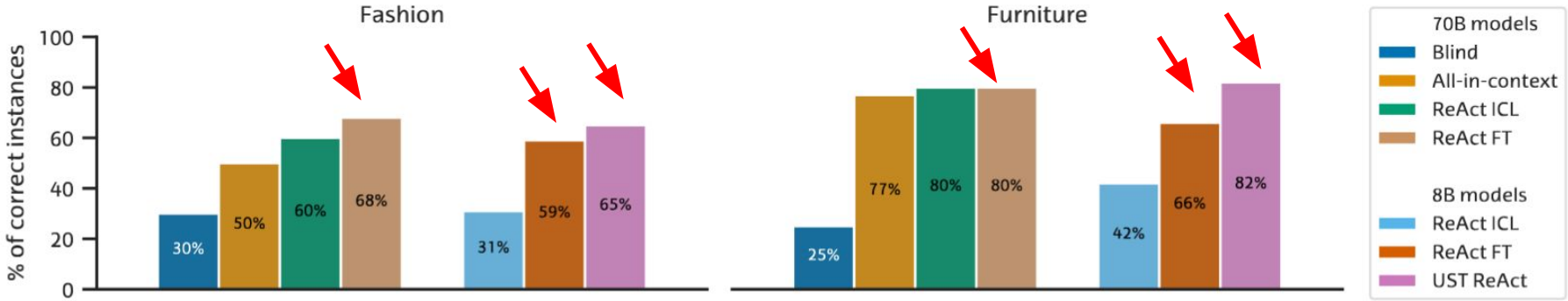
- ReAct outperforms a chatbot having access to all data in context (All-in-context)



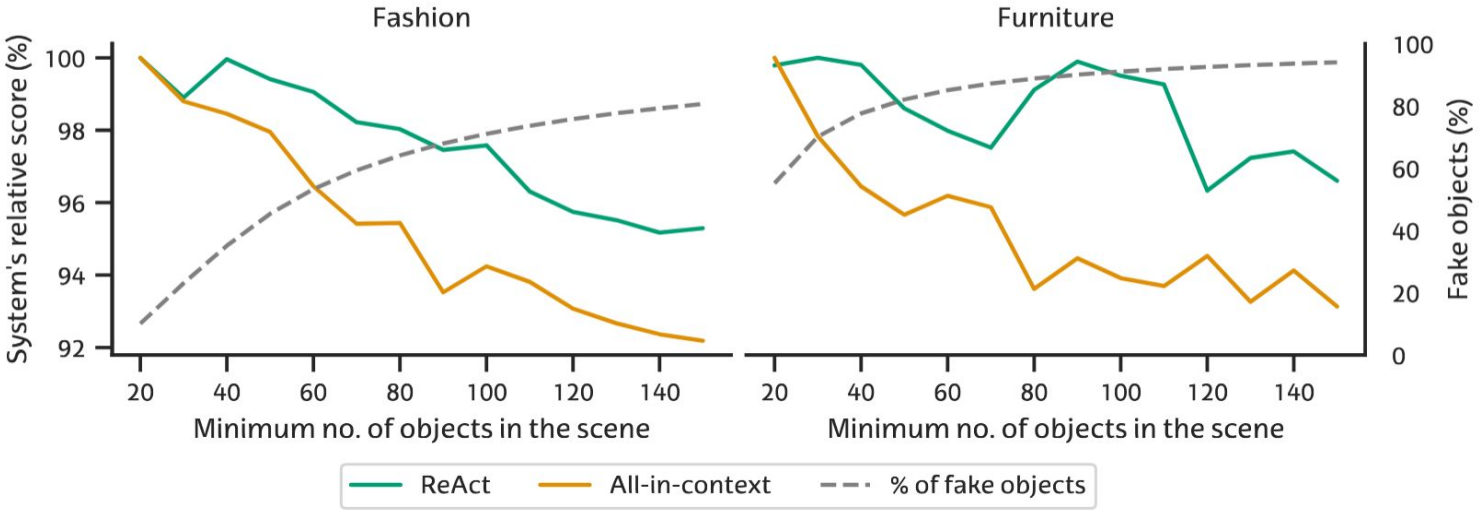
Applying ReAct to SIMMC2.1

- Develop an agentic chatbot that acts as a salesperson.
- **Manual Evaluation**: Answers must be factually correct and relevant to the question.

- ReAct outperforms a chatbot having access to all data in context (All-in-context)
- Fine-tuning on correct trajectories improve results (ReAct FT, UST ReAct)



Performance when Many Objects Co-occur

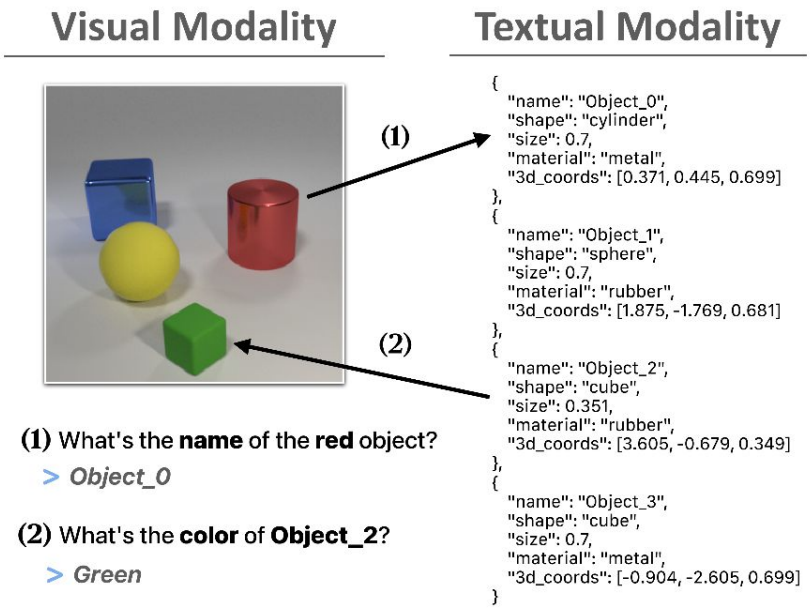


- ReAct performs betters when there are many objects in the environment (metadata)

Limitations of Current Multimodal LLMs

Visual-Language models struggle to link cross-modal entities

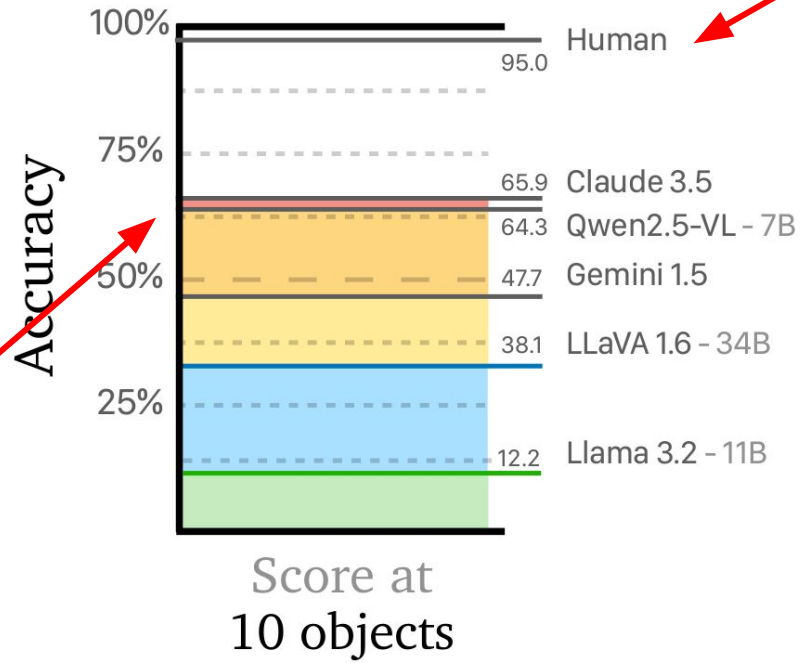
- MATE: Dataset for benchmarking cross-modal entity linking
- Trivial task for humans but requires truly understanding both modalities
  - **Visual search:** what is the red object?
  - **Identifying linking attributes:** what other attributes distinguish the object?
  - **Textual search:** which object in the textual modality has these distinguishing attributes?



MATE: A Benchmark to Evaluate Cross-Modal Entity Linking

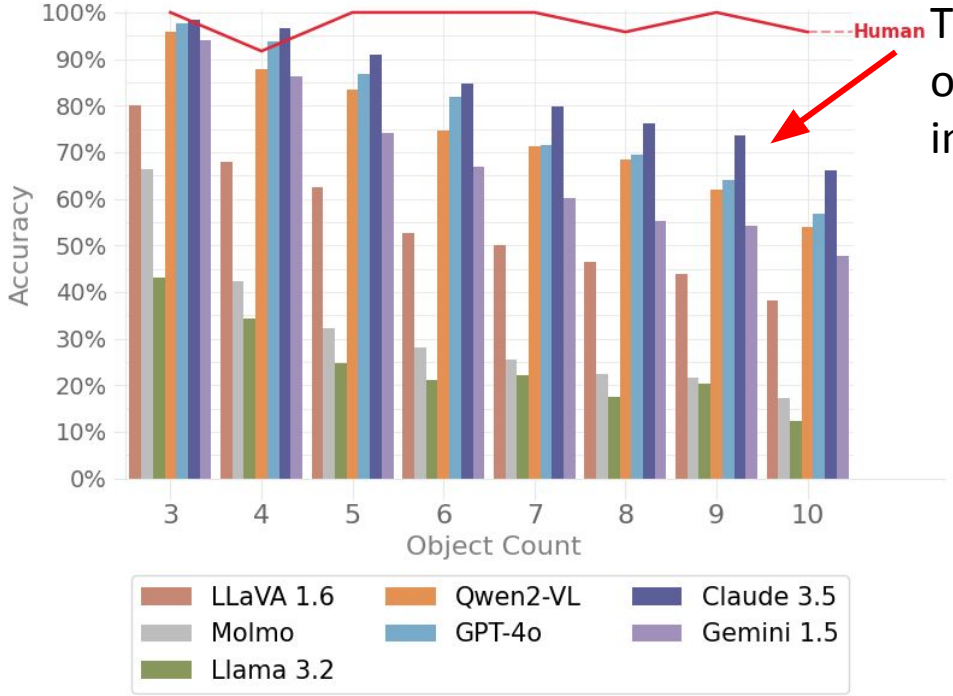
- We introduce MATE, a benchmark specifically designed to isolate and evaluate the ability of VLMs to perform cross-modal entity linking.

Trivial for humans



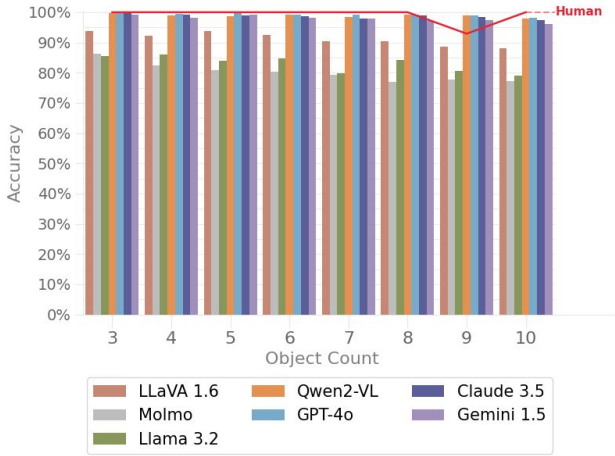
Current SOTA models struggle completing the task.

The task is easy, cross-modality is the issue



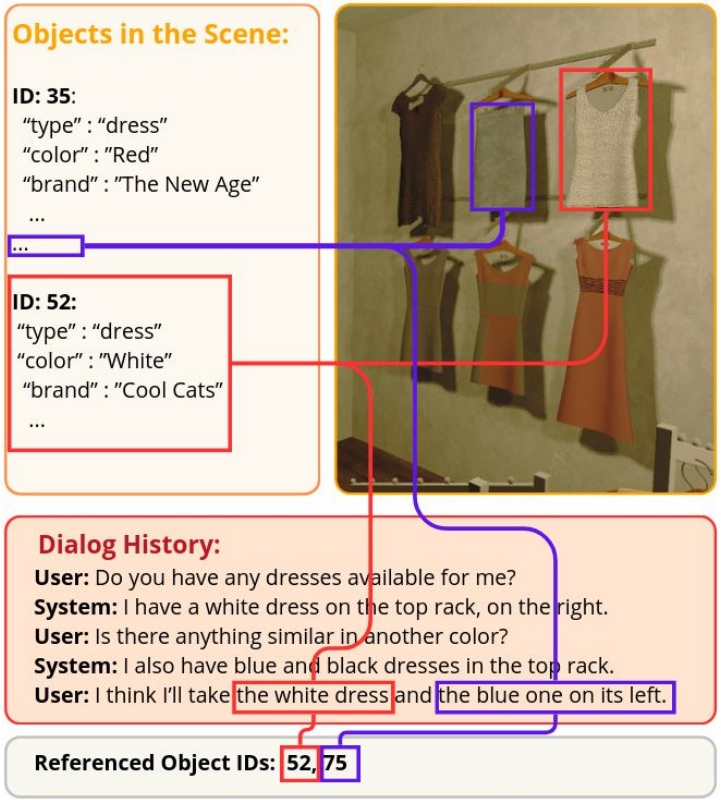
They fail as the number of objects in the scene increases.

Not the case in **unimodal**: only text or only image



Can we improve linkage across modalities?

- **Task**: Multimodal Coreference Resolution



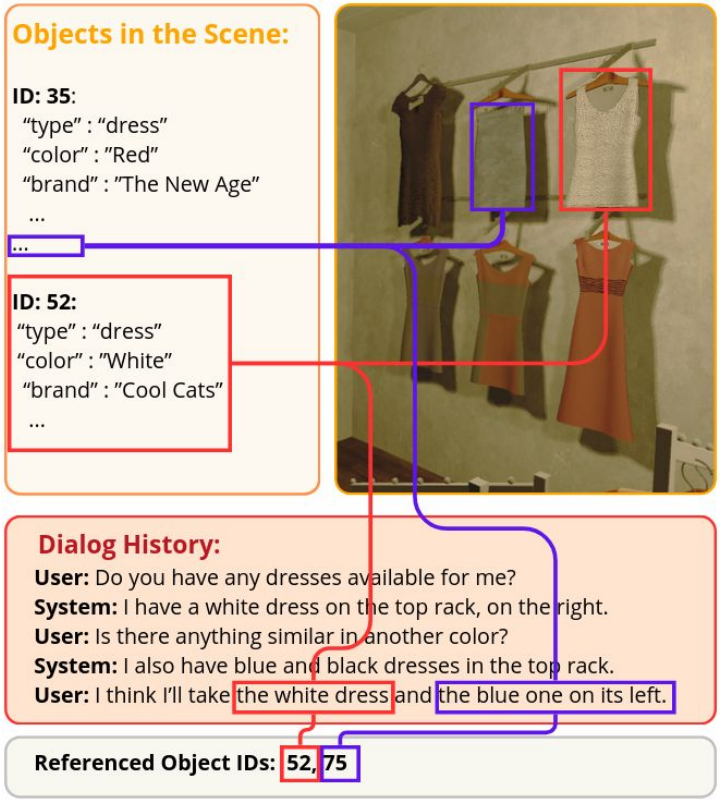
Given the image of the **scene**, object **metadata**, and **dialogue** history, the model must identify the object references in the last utterance of the user (e.g. "**the white dress**" with **object id 52**).

Can we improve linkage across modalities?

- **Task**: Multimodal Coreference Resolution

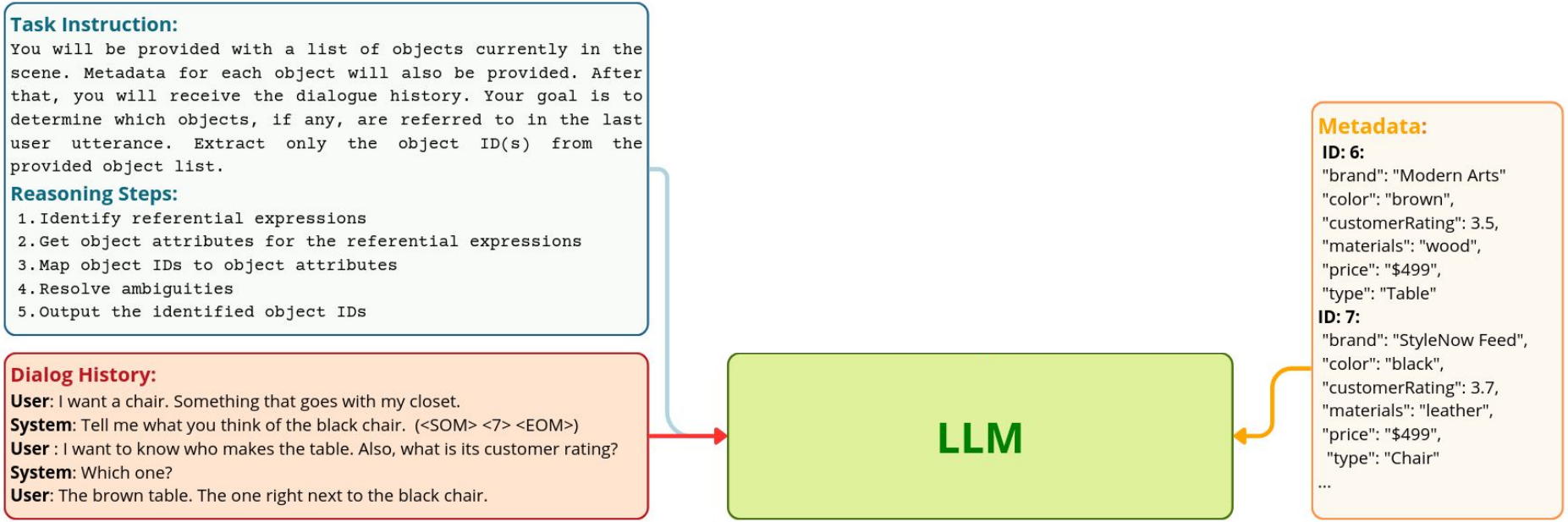
We show that:

- **Test-time reasoning** enables LLMs to reason over detailed object metadata and dialogue history to improve **coreference resolution**.
- **LLMs can generate step-by-step reasoning** that effectively **align** dialogue context with objects present in the scene.



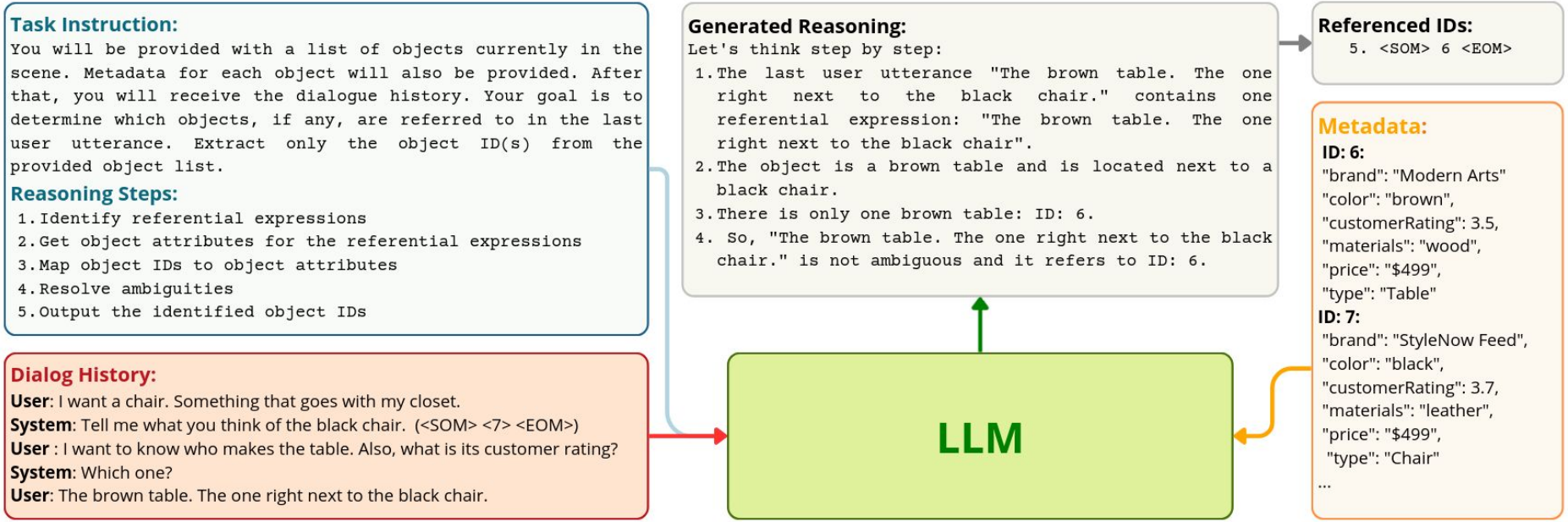
Given the image of the **scene**, object **metadata**, and **dialogue** history, the model must identify the object references in the last utterance of the user (e.g. "**the white dress**" with **object id 52**).

Reasoning over Object Descriptions



- The LLM receives **task instruction** with **reasoning steps**, object descriptions (**metadata**) and the previous dialogue and object references (**dialogue history**).

Reasoning over Object Descriptions



- The LLM receives **task instruction** with **reasoning steps**, object descriptions (**metadata**) and the previous dialogue and object references (**dialogue history**).
- The LLM generates some **reasoning** that delivers the **referenced object IDs** by the user (<SOM>6<EOM>)

Improves Coreference Resolution in Task-Based Dialogue Systems

	Model	Size	Precision	Recall	F1 Score
	Random		2.68	49.87	5.09
Zero-shot	Qwen3	4B	**29.01**	63.91	39.90
	Gemma	7B	16.93	52.74	25.64
	Mistral	7B	25.70	66.62	37.09
	Qwen2.5	7B	24.14	71.66	36.12
	Llama3.1	8B	27.74	63.03	38.53
	DeepSeek-R1	8B	22.98	66.77	34.19
	Llama3.3	70B	26.61	**80.10**	**39.95**
Few-shot	Qwen3	4B	39.13(+10.12)	59.15(-4.76)	47.10(+7.2)
	Gemma	7B	19.58(+2.65)	41.38(-11.36)	26.58(+0.94)
	Mistral	7B	29.14(+3.44)	62.65(-3.97)	39.77(+2.68)
	Qwen2.5	7B	28.84(+4.7)	70.05(-1.61)	40.86(+4.74)
	Llama3.1	8B	30.34(+2.6)	68.39(+5.36)	42.03(+3.5)
	DeepSeek-R1	8B	25.31(+2.33)	67.17(+0.4)	36.77(+2.58)
	Llama3.3	70B	**39.32**(+12.71)	**73.12**(-6.98)	**51.14**(+14.34)

Few-shot improvement. Adding only 3 examples in the prompt shows consistent improvements of the F1 score.

Improves Coreference Resolution in Task-Based Dialogue Systems

	Model	Size	Precision	Recall	F1 Score
	Random		2.68	49.87	5.09
Zero-shot	Qwen3	4B	**29.01**	63.91	39.90
	Gemma	7B	16.93	52.74	25.64
	Mistral	7B	25.70	66.62	37.09
	Qwen2.5	7B	24.14	71.66	36.12
	Llama3.1	8B	27.74	63.03	38.53
	DeepSeek-R1	8B	22.98	66.77	34.19
	Llama3.3	70B	26.61	**80.10**	**39.95**
Few-shot	Qwen3	4B	39.13(+10.12)	59.15(-4.76)	47.10(+7.2)
	Gemma	7B	19.58(+2.65)	41.38(-11.36)	26.58(+0.94)
	Mistral	7B	29.14(+3.44)	62.65(-3.97)	39.77(+2.68)
	Qwen2.5	7B	28.84(+4.7)	70.05(-1.61)	40.86(+4.74)
	Llama3.1	8B	30.34(+2.6)	68.39(+5.36)	42.03(+3.5)
	DeepSeek-R1	8B	25.31(+2.33)	67.17(+0.4)	36.77(+2.58)
	Llama3.3	70B	**39.32**(+12.71)	**73.12**(-6.98)	**51.14**(+14.34)
Reasoning	Qwen3	4B	**59.98**(+20.85)	45.21(-13.94)	51.55(+4.45)
	Gemma	7B	37.86(+18.28)	43.24(+1.86)	40.37(+13.79)
	Mistral	7B	39.82(+10.68)	52.93(-9.72)	45.45(+5.68)
	Qwen2.5	7B	49.90(+21.06)	57.32(-12.73)	53.35(+12.49)
	Llama3.1	8B	52.16(+21.82)	59.23(-9.16)	55.47(+13.44)
	DeepSeek-R1	8B	53.50(+28.19)	43.92(-23.25)	48.24(+11.47)
	Llama3.3	70B	58.73(+19.41)	**61.37**(-11.75)	**60.02**(+8.88)

Few-shot improvement. Adding only 3 examples in the prompt shows consistent improvements of the F1 score.

Reasoning vs few-shot. The application of reasoning on top of few-shot learning also yields positive results in terms of F1 scores (19 points of average improvement).

Improves Coreference Resolution in Task-Based Dialogue Systems

	Model	Size	Precision	Recall	F1 Score
	Random		2.68	49.87	5.09
Zero-shot	Qwen3	4B	**29.01**	63.91	39.90
	Gemma	7B	16.93	52.74	25.64
	Mistral	7B	25.70	66.62	37.09
	Qwen2.5	7B	24.14	71.66	36.12
	Llama3.1	8B	27.74	63.03	38.53
	DeepSeek-R1	8B	22.98	66.77	34.19
	Llama3.3	70B	26.61	**80.10**	**39.95**
Few-shot	Qwen3	4B	39.13(+10.12)	59.15(-4.76)	47.10(+7.2)
	Gemma	7B	19.58(+2.65)	41.38(-11.36)	26.58(+0.94)
	Mistral	7B	29.14(+3.44)	62.65(-3.97)	39.77(+2.68)
	Qwen2.5	7B	28.84(+4.7)	70.05(-1.61)	40.86(+4.74)
	Llama3.1	8B	30.34(+2.6)	68.39(+5.36)	42.03(+3.5)
	DeepSeek-R1	8B	25.31(+2.33)	67.17(+0.4)	36.77(+2.58)
	Llama3.3	70B	**39.32**(+12.71)	**73.12**(-6.98)	**51.14**(+14.34)
Reasoning	Qwen3	4B	**59.98**(+20.85)	45.21(-13.94)	51.55(+4.45)
	Gemma	7B	37.86(+18.28)	43.24(+1.86)	40.37(+13.79)
	Mistral	7B	39.82(+10.68)	52.93(-9.72)	45.45(+5.68)
	Qwen2.5	7B	49.90(+21.06)	57.32(-12.73)	53.35(+12.49)
	Llama3.1	8B	52.16(+21.82)	59.23(-9.16)	55.47(+13.44)
	DeepSeek-R1	8B	53.50(+28.19)	43.92(-23.25)	48.24(+11.47)
	Llama3.3	70B	58.73(+19.41)	**61.37**(-11.75)	**60.02**(+8.88)

Few-shot improvement. Adding only 3 examples in the prompt shows consistent improvements of the F1 score.

Reasoning vs few-shot. The application of reasoning on top of few-shot learning also yields positive results in terms of F1 scores (19 points of average improvement).

Model size. Results suggest that model size improve the ability to correctly perform reasoning steps.

Conclusions

Conclusions

- We show various showcases where LLMs can assist in virtual environments
- Cross-modal linkage is not trivial, but it is required for effective grounding.
- Current models **struggle** to link cross-modal entities.

- To effectively assist:
  - **Grounding techniques** are key to make it aware of the surrounding environment.
  - **Test-time reasoning** and advanced prompting can alleviate the linking problem.
  - If data is available **smart fine-tuning** could help obtaining better generalization.

Thanks!
Questions?