Talk Less, Interact Better: Evaluating In-context Conversational Adaptation in Multimodal LLMs

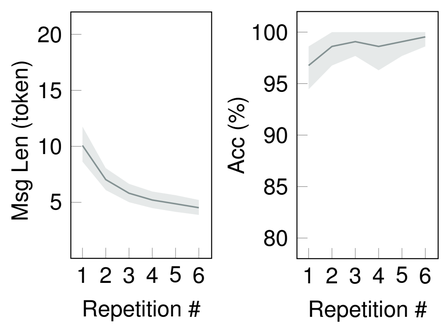
Yilun Hua and Yoav Artzi

Paper & Code

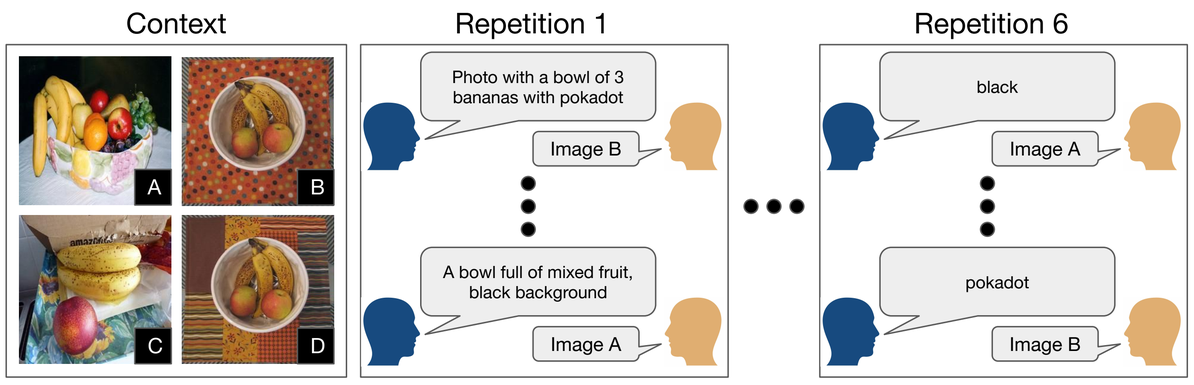
Humans communicate with increasing efficiency through repeated interactions

Adapt to use more concise, conventionalized language (ad-hoc conventions)

- Convergence
- Stability
- Understand the partner better



[Hawkins et al., 2020]



ICCA - An Automated Evaluation Framework (**I**n-**c**ontext **C**onversational **A**daptation)

- Can MLLM agents spontaneously adapt to use more efficient language?
- Can MLLM agents better understand a partner who is adopting more efficient language over time?

Why do we expect MLLMs to adapt like humans?

They are trained on large-scale human data, where efficiency adaptation is common.

Why is adapting like humans important for MLLMs?

Adaptation is critical for efficient and natural conversations.

ICCA Framework

Overview

At every trial

- Preprocesses the interaction context into a query prompt
- Queries the model; computes the feedback

Preprocessor is customizable

- Supports various interaction variants
- Vary the instructions, ways to present the images

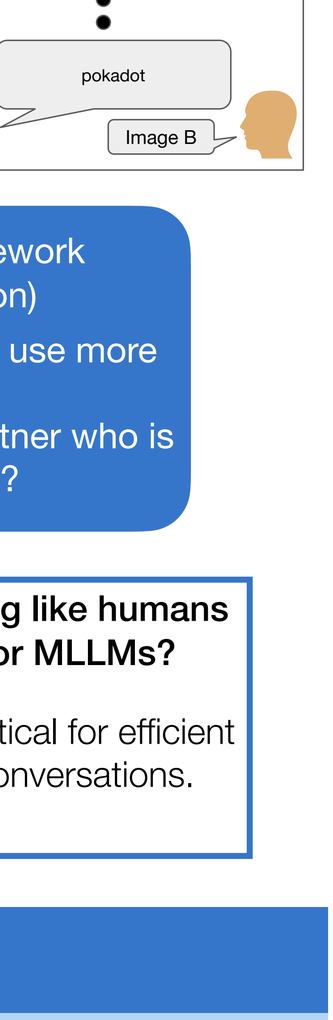
Automated Eval - Simulated Interlocutor

Model-as-listener eval with deterministic speaker

- Messages from human interactions
- Predetermined, realistic trajectories of efficiency

Model-as-speaker eval with GPT4 listener:

- High performance, similar to humans



Model-as-Speaker

Design interaction variants by changing the instruction

Standard

```
[System] ... Each time, generate a message to tell the listener which image
is the target ...

Image A: <image>  Image B: <image>  Image C: <image>  Image D: <image>

Trial 1, the target is Image B.
[Speaker] Message: Photo with a bowl of 3 bananas with a pokadot background
[System] The listener correctly answered Image B.

Trial 2, the target is Image A.
[Speaker] Message: A bowl of mixed fruit, black background
```

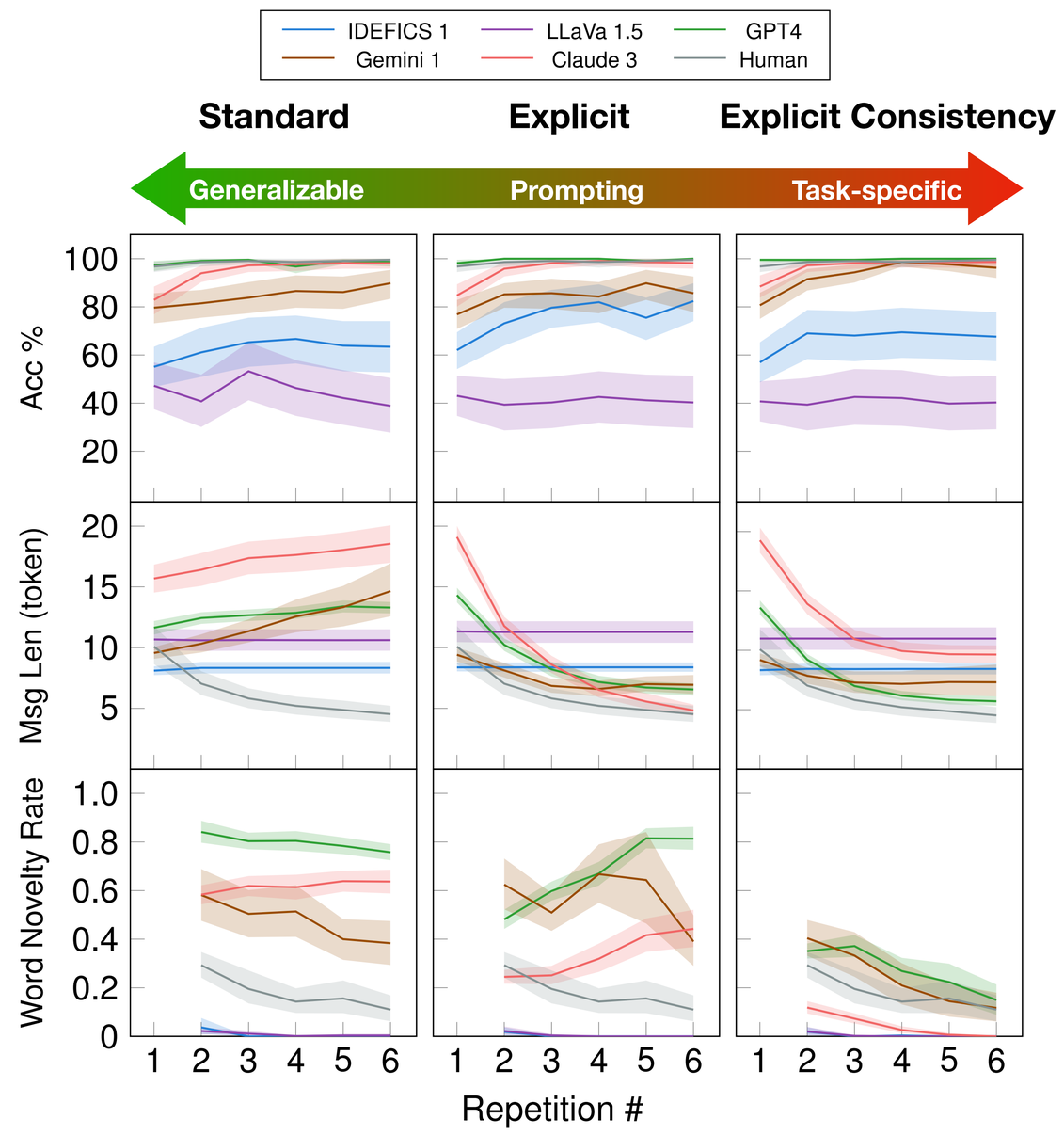
⋮ Testing increasingly explicit instructions

Explicit

```
[System] ... make your messages shorter and shorter every trial ...
```

Explicit Consistency

```
[System] ... shorten the messages by extracting salient tokens from the previous
messages; keep using the same message if it cannot be shortened further...
```



Human Messages

1. dirty truck going towards the bridge
2. dirty truck going to bridge
3. how is this even guessing, it's so easy, dirty truck going to bridge
4. dirty truck
5. dirty truck
6. dirty truck

GPT4 Messages Explicit Instruction

1. construction bridge with a concrete mixer truck underneath
2. bridge construction with a white truck
3. bridge work, dirty truck below
4. bridge construction
5. overpass work
6. concrete mixer

More in the paper

- Other interaction variants
- Model-as-Listener Experiments: performance depends on the interaction complexity
- Prompting for Gricean behaviors is insufficient for adaptation
- Models exploit shortcuts to show convention formation without visual grounding

Conclusion

- MLLMs do not spontaneously show ad-hoc adaptations for efficient communication
- MLLMs do not adequately model the ad-hoc adaptation present in their training data