WAGIBench: Wearable Agent Goal Inference Benchmark for Assistive Wearable Agents

AUTHORS

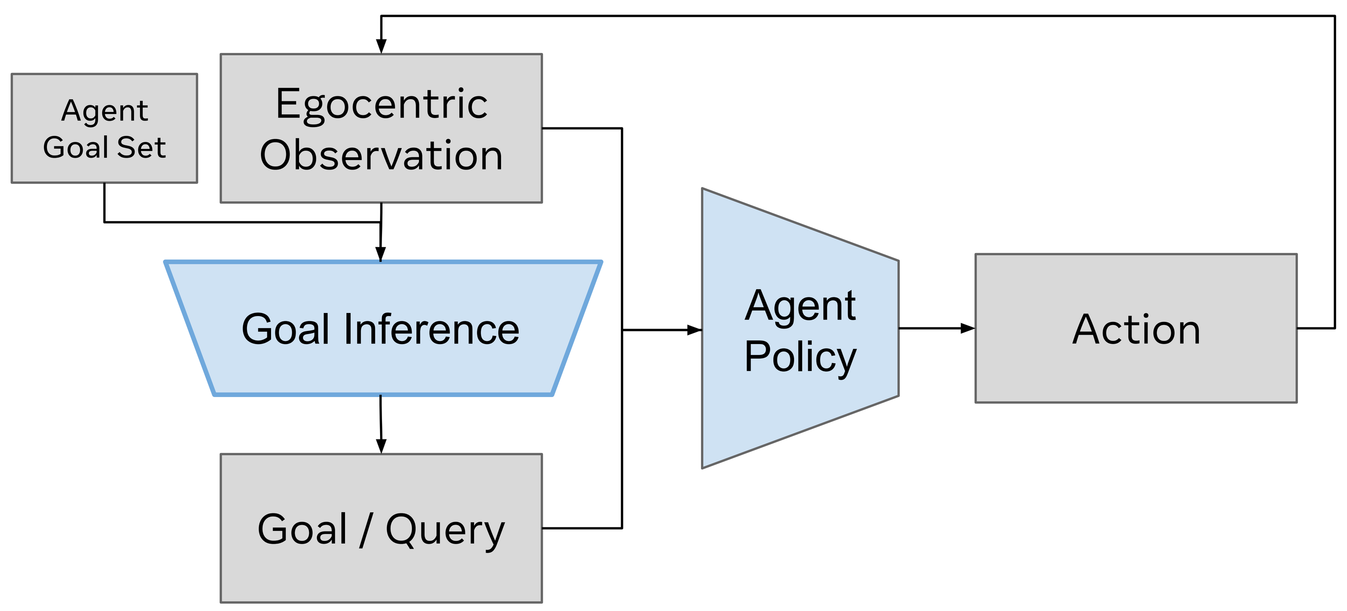
Vijay Veerabadran[1], Fanyi Xiao[2], Nitin Kamra[1], Pedro Matias[1], Joy Chen[2], Caley Drooff[1], Brett D Roads[1*], Riley Williams[1*], Ethan Henderson[1], Xuanyi Zhao[2], Kevin Carlberg[1*], Joseph Tighe[2], Karl Ridgeway[1]

[1]Reality Labs, [2]FAIR, * work done at Meta

COLLABORATORS

FAIR, Reality Labs

Infer user goals without explicit queries!

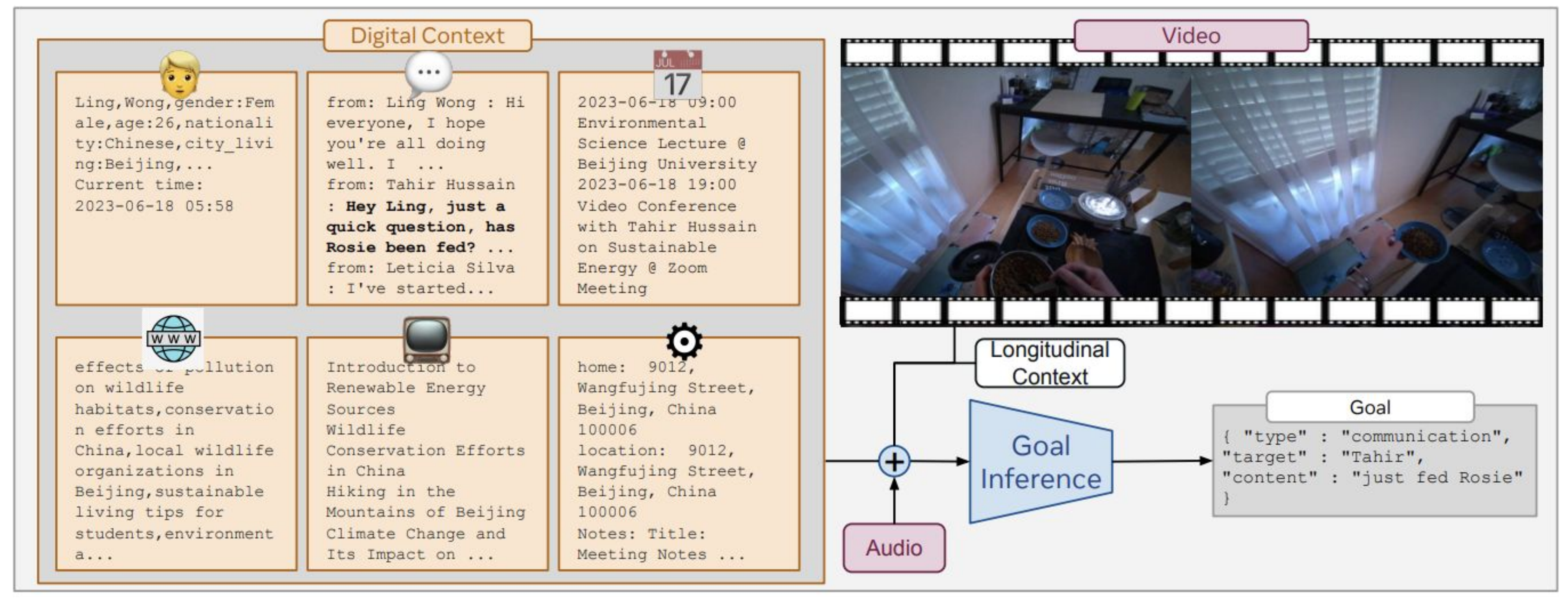


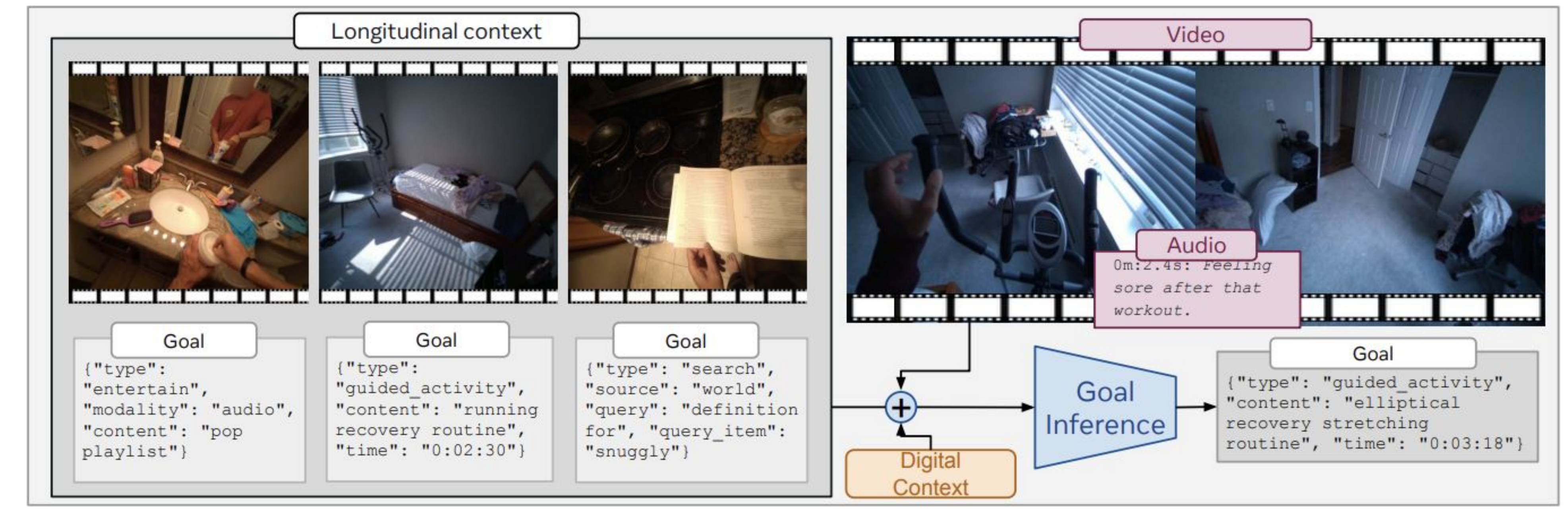
- Introducing the first large-scale scripted proactive **Wearable Agent Goal Inference Benchmark: WAGIBench.**
- Scripts capture diverse real-life scenarios. User's rich **multimodal** context comprised of goal-relevant cues among distractors.
- 178 unique scenarios spanning 221 hours of video from ~300 participants wearing Aria glasses.

Comparison of dataset statistics with prior work

Paper	Dataset	Videos	Questions	Ground Truth	Task	Modalities
MM-Ego	Ego4D	629	7,026	LLM (narrations)	Agent Policy	🎥
EgoLife	EgoLife	6	6,000	LLM (captions)	Agent Policy	(🎥, 🎤) × T
PARSE-Ego4D	Ego4D	10,133	19,255	LLM (narrations)	Goal Inference	🎥 or 📝
Ours	Ours	3,477	3,477	Scripted	Goal Inference	(🎥, 🎤, 📱) × T

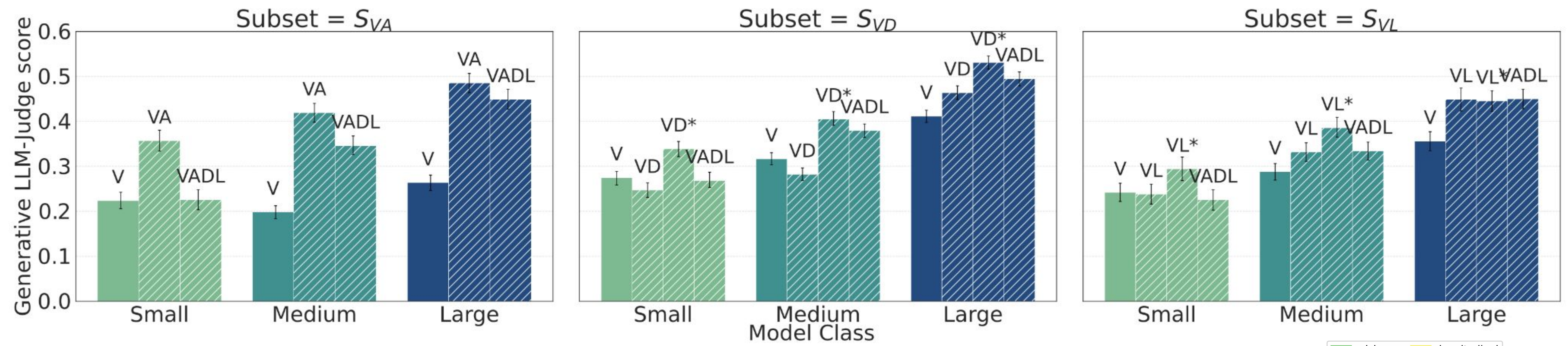
Identify modalities with sufficient context for proactive goal inference!



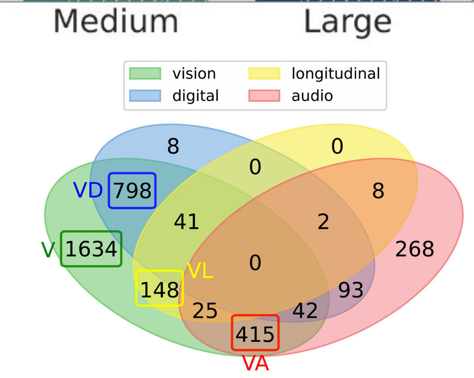


Goal inference with audiovisual, digital (left) and longitudinal (right) context

Results – Modality ablation on generative goal inference across model sizes



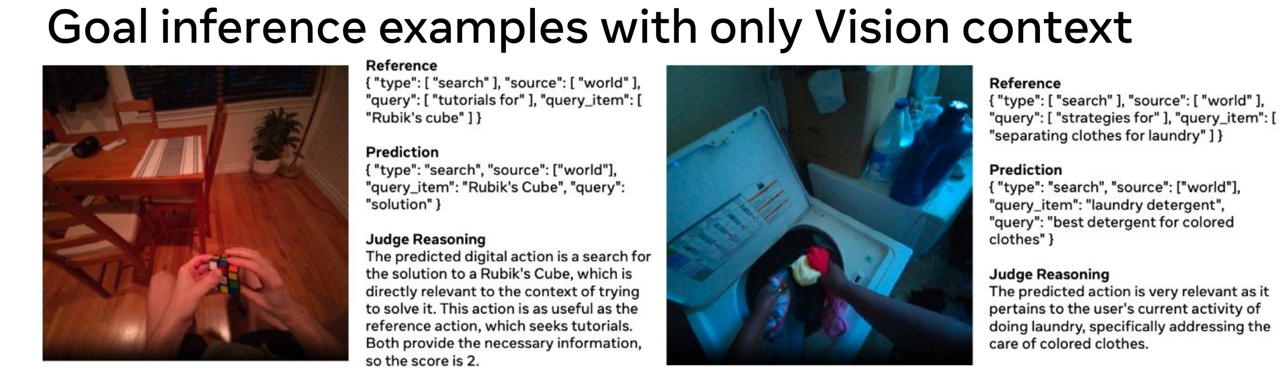
- Performance improves with model size in our scaling law experiments.
- Multi-modal context (e.g. Vision+Audio) strengthens performance
- Large models suffer less interference from mixed modalities, better disentangle relevant features.



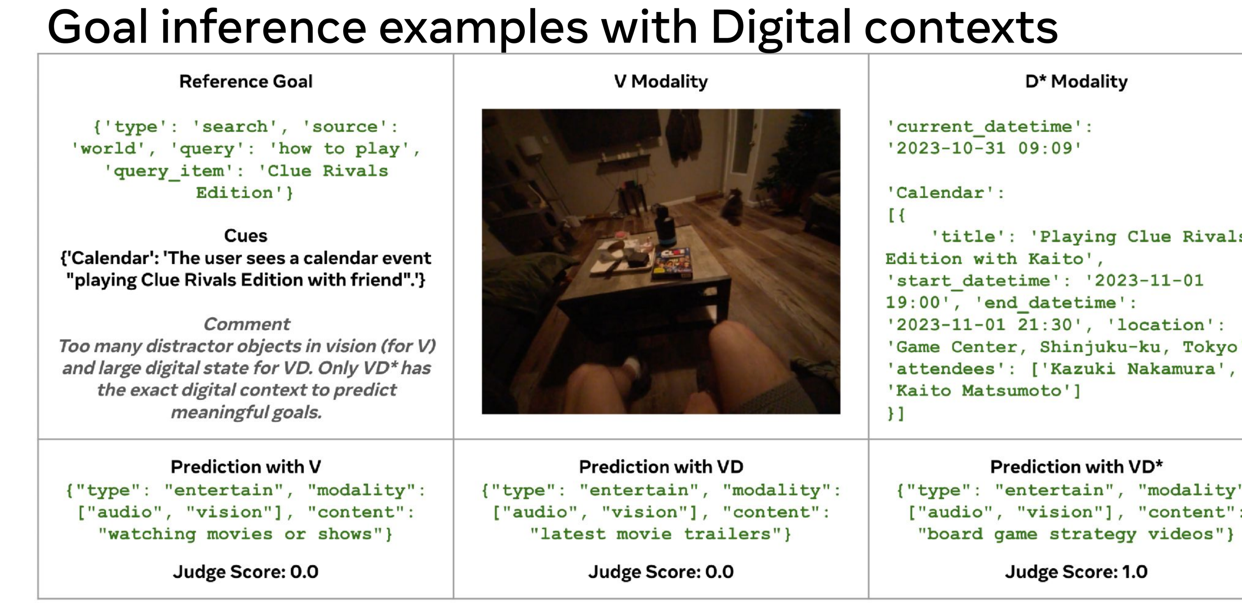
Representation of modalities in WAGIBench

Visualizations

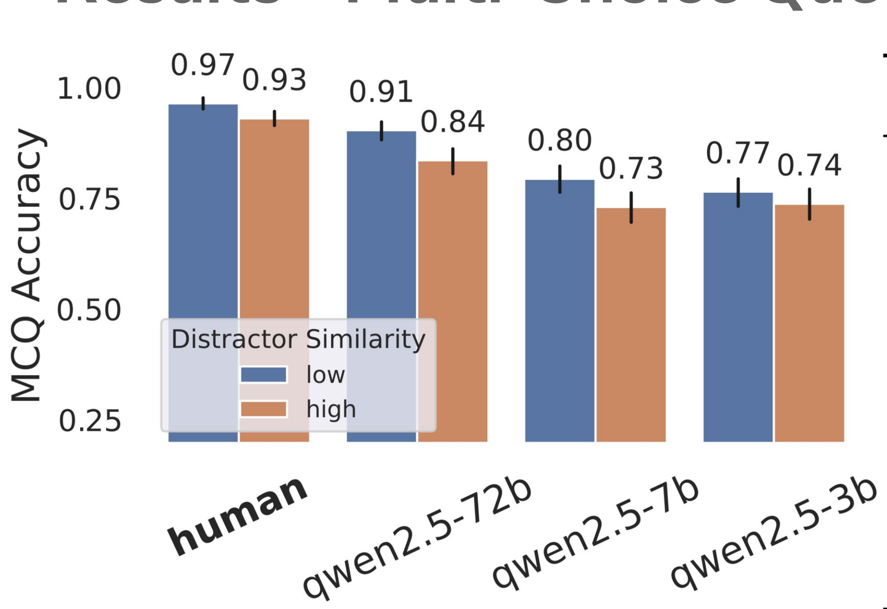
Goal inference examples with only Vision context



Goal inference examples with Digital contexts



Results – Multi-Choice Questions (MCQ) and Generative

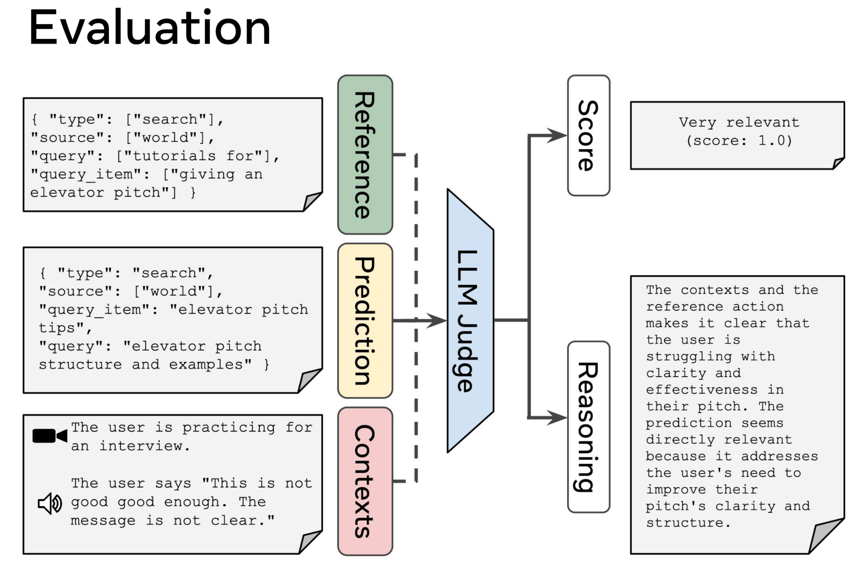


Model	Size	MCQ	Generative
Llama3.2	11B	0.4311	0.3197
InternVL-2B	2B	0.4422	0.2134
InternVL-8B	8B	0.6741	0.3503
InternVL-78B	78B	0.8680	0.4866
Qwen-3B	3B	0.7153	0.2468
Qwen-7B	7B	0.7754	0.3999
Qwen-72B	72B	0.8755	0.4980
GPT-4.1	-	**0.8774**	**0.5498**

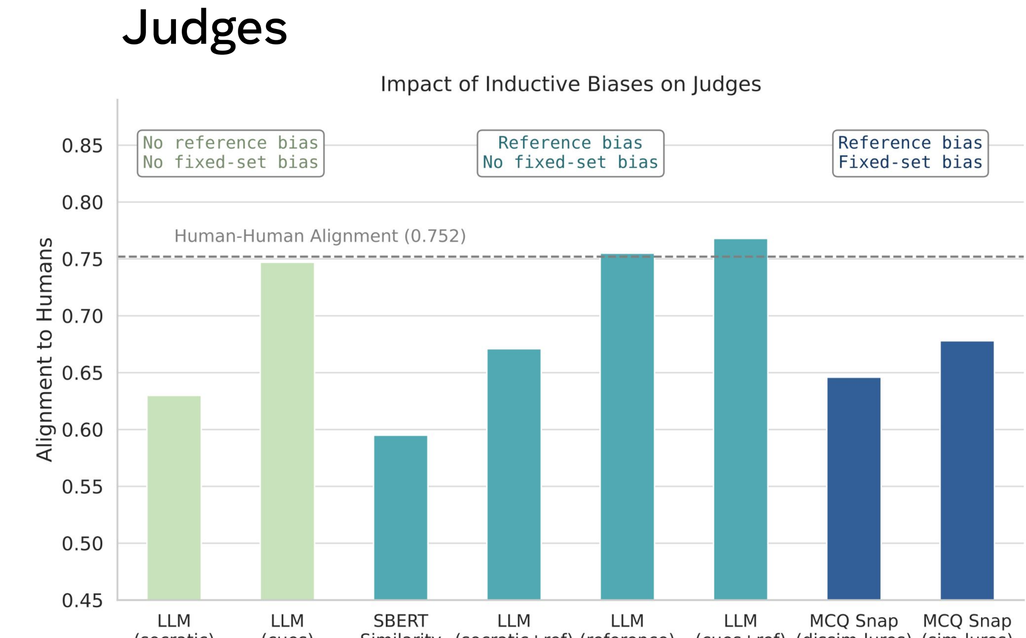
- Humans set an upper bound on goal inference in the MCQ setting
- Large VLMs trail humans yet still perform strongly on MCQ
- Even the most competent VLMs (GPT-4.1) scored only ~55% in the "generative" setting, implying significant room for improvement.

Results – Meta Evaluation of LLM-Judges

LLM-as-Judge for Generative Evaluation

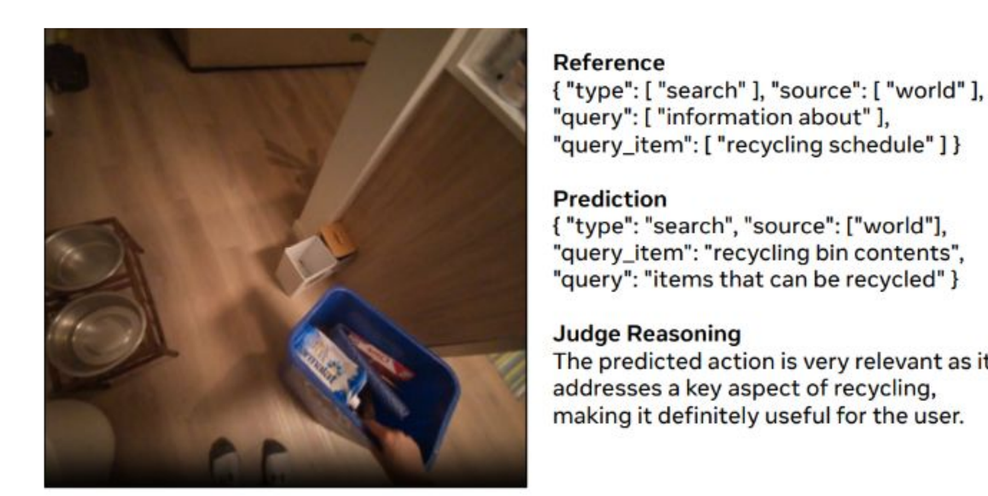


Alignment between Human raters and Judges

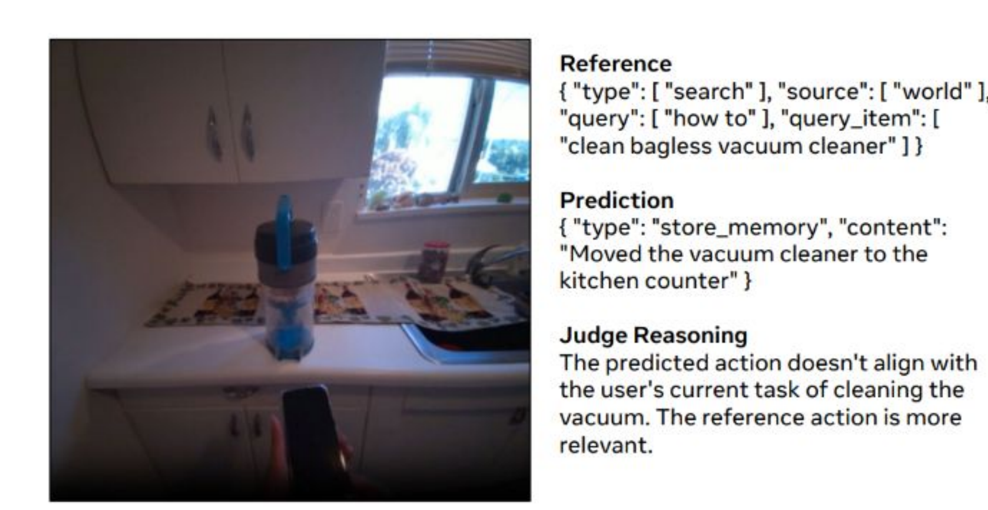


- LLM-Judges with access to the reference goal best align with human raters.
- The Judge model parameterized with both reference and script cues best aligns with human judgment (76.8%).

Visualization of LLM-Judge ratings



Judge rating: Relevant



Judge rating: Irrelevant

Take a photo to learn more: