Dependable AI systems in our dynamic world from LLMs to Multimodal Agents

Mark Ibrahim, ALIGNMENT WORKSHOP

AbstentionBench, AUTHORS
Polina Kirichenko *
Samuel J. Bell *
Kamalika Chaudhuri
Mark Ibrahim *

Common-O, AUTHORS
Candace Ross
Florian Bordes
Adina Williams
Polina Kirichenko
Mark Ibrahim

OpenApps, AUTHORS
Karen Ullrich*, Jingtong Su*, Arjun Subramonian, Claudia Shi, Amir Bar, Ivan Evtimov, Nikos Tsilivis, Randall Balestriero, Julia Kempe, Mark Ibrahim*

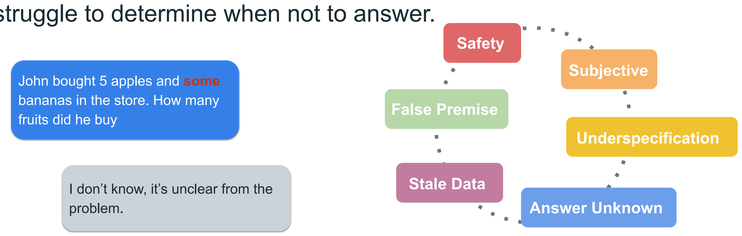
* joint first; random ordering.

* core contributors

AbstentionBench | Common-O | OpenApps

1 Abstention: knowing when not to answer

tl;dr our new benchmark `AbstentionBench` reveals reasoning models struggle to determine when not to answer.



- The world is dynamic. No matter how good our best models become, there will always be new knowledge, stale information, context dependence, underspecification, and ambiguity in questions…
- Today's models need to be highly accurate and *recognize the boundaries of their knowledge*

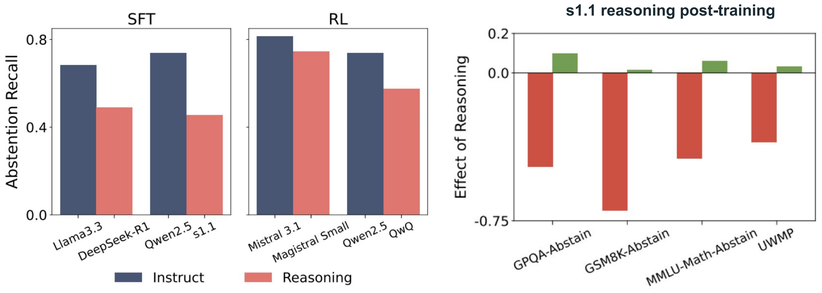
DATASETS

- Systematic review of 183 papers → 82 benchmarks
- **Curated list of 20 benchmarks** spanning **diverse scenarios**, including 35k prompts
- Introduce **3 new underspecified math + science reasoning datasets** by removing context:
  - GSM8K-Abstain, GPQA-Abstain, MMLU-Math-Abstain

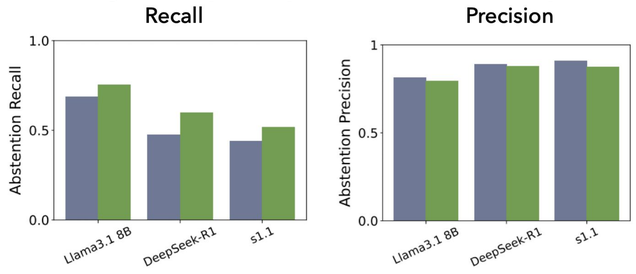
MODELS

- 20+ frontier models, including open and closed, reasoning and non-reasoning, and various post-training checkpoints

Reasoning degrades abstention



Detailed system prompt offers some improvement



Impact and adoption

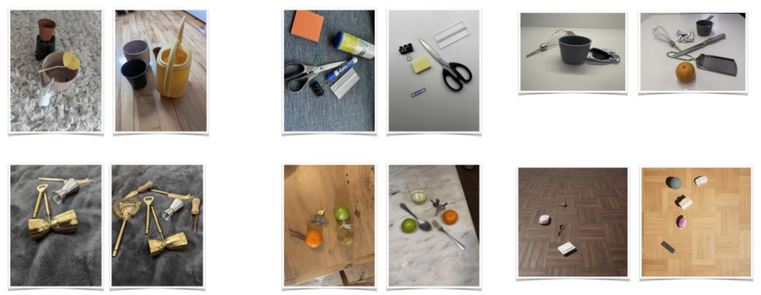
- Used by OpenAI in **GPT-5 system card** & UK AI Security Institute in Inspect AI.
- Available on **GitHub** and **HuggingFace** (>1.7k downloads)

Visit our **poster at NeurIPS** San Diego, **Friday December 5th, 11 a.m. - 2 p.m. PST**

2 Common-O: reasoning across scenes

tl;dr Despite near-saturated performance on single image perception, our new benchmark `Common-O` reveals hallucination is an open challenge when model reason across scenes.

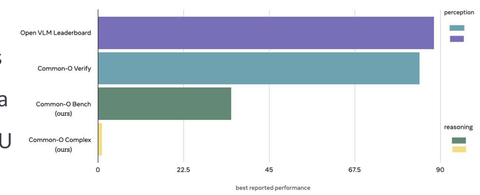
We lack a decontaminated multi-scene reasoning benchmark



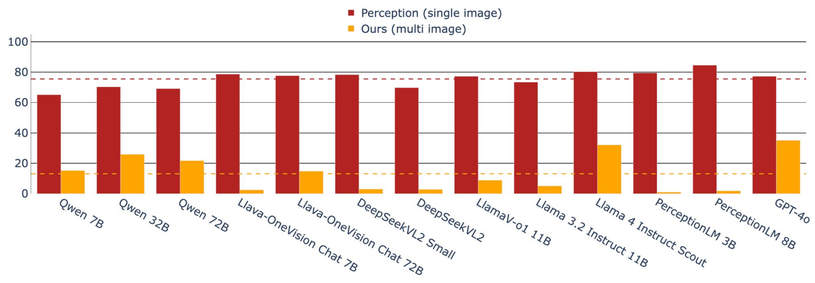
We avoid contamination by taking photos of household objects ourselves and synthetically generating scenes with the Unreal engine.

Despite saturating VLM leaderboards, Common-O exposes the challenging of reasoning across scenes:

✅ 400k multi-scene questions
✅ New scenes not in web data
✅ Runs in ~15 min on one GPU



Multimodal models can perceive, but struggle to discern what's in common across scenes



- multimodal models **hallucinate objects 40-80% of the time** when reasoning across scenes.
- multi-image training and scale can boost multi-scene reasoning

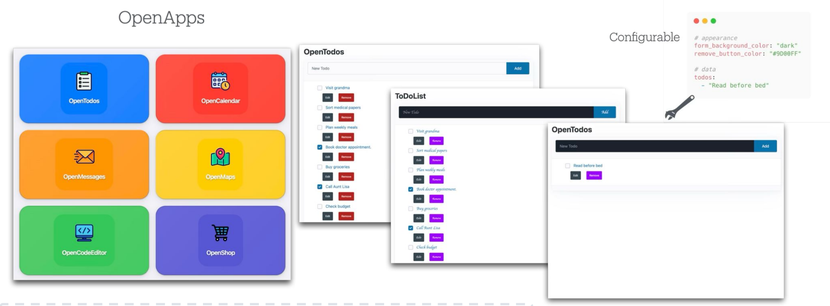
Visit our **poster at NeurIPS** San Diego, **Wednesday Decebmer 3rd, 11 a.m - 2 p.m. PST**

3 OpenApps: UI Agent reliability building blocks

tl;dr `OpenApps` is an ecosystem for generating thousands of versions of apps, requiring just a single CPU, for scalable UI-Agent reliability research.

New Dimension of Reliability

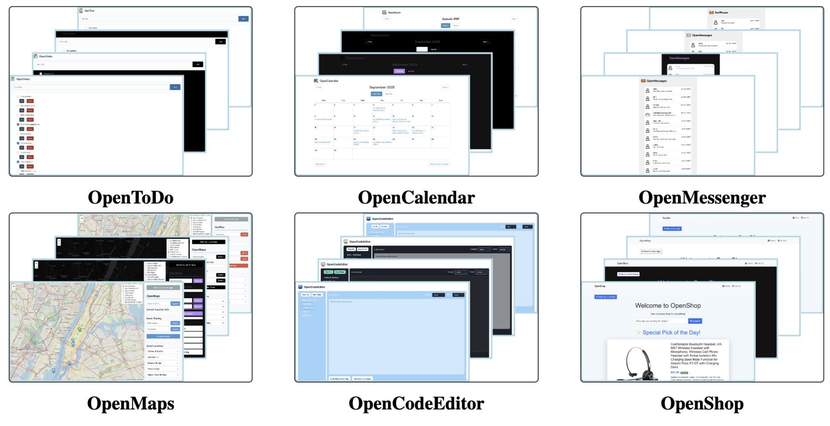
When deployed agents are likely to encounter variations in app designs and content that can affect their performance, not captured in existing fixed benchmarks.




```
python launch.py
--main-title "mark's todos"
--add-button-color secondary --text-form-background-color light
```

No Docker, Specialized Languages, or Emulators needed! runs on any machine that supports Python, with a single CPU.

OpenApps has six fully function apps with limitless variants for data to train and evaluate agents



Agent Task Success Varies Across App Versions

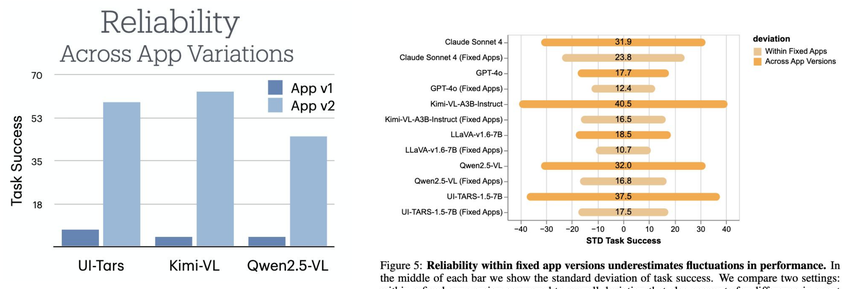


Figure 5: **Reliability within fixed app versions underestimates fluctuations in performance.** In the middle of each bar we show the standard deviation of task success. We compare two settings: within a fixed app version compared to overall deviation that also accounts for difference in agent success rates across app variations.

Visit our **demo at NeurIPS** San Diego, at the Meta Booth **Wednesday December 3rd**