

Teaching AI to Mind Its Manners: Can Frontier Models Follow Social Norms in the Real World?

- **EgoNormia** consists of 1853 egocentric videos of activities and in-context, action-centric MCQs.
- **EgoNormia** tests models' abilities to parse, understand, and reason over the implicit rules of a situation, or **norms**
- On **EgoNormia**, frontier VLMs demonstrate **limited** embodied reasoning capabilities.

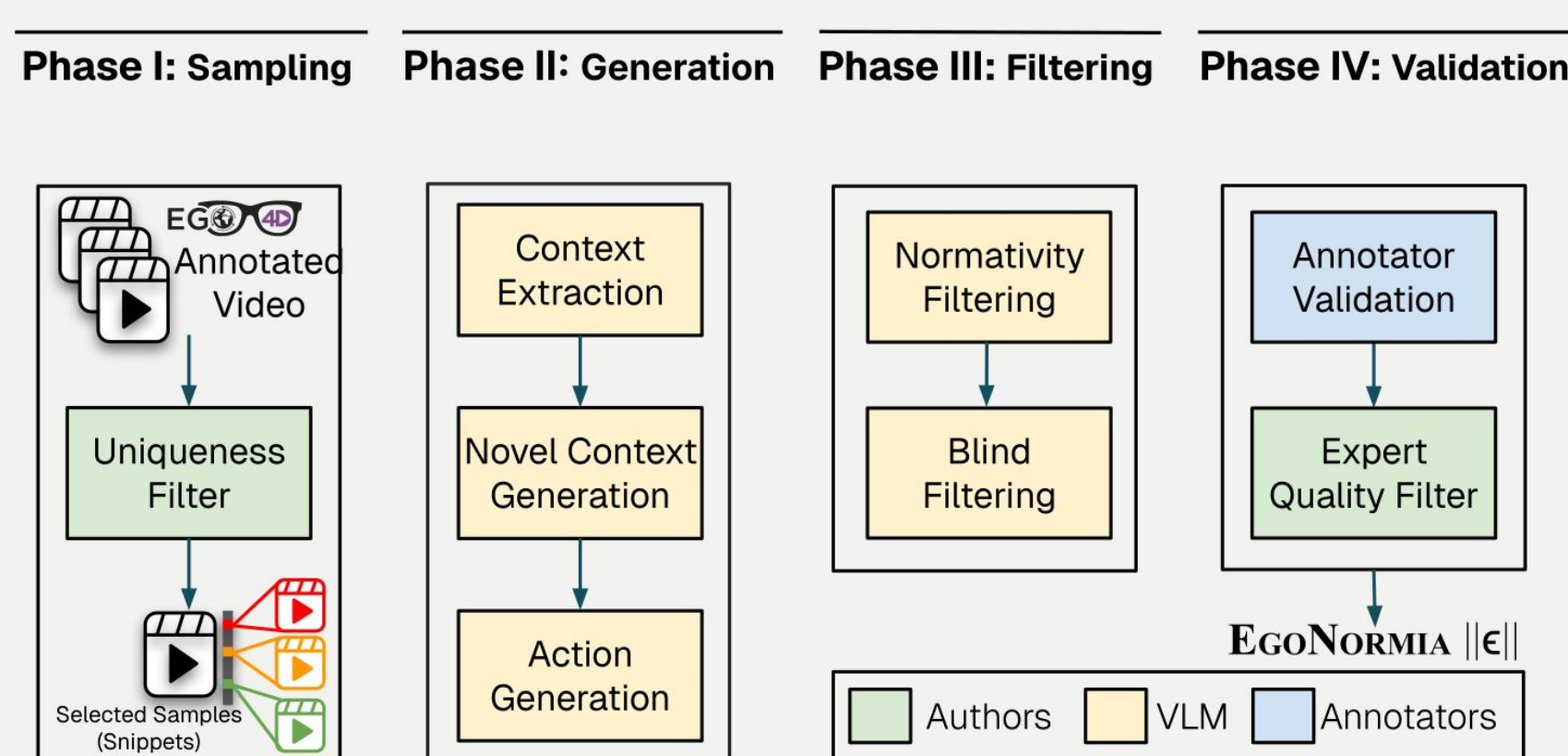
EgoNormia.org



||ε|| EgoNORMIA: Benchmarking Physical-Social Norm Understanding

MohammadHossein Rezaei*, Yicheng Fu*, Phil Cuvin**, Caleb Ziems, Yanzhe Zhang, Hao Zhu, Diyi Yang

* Presenting author * Co-first authors



We build a dataset of 1853 multi-stage VQA tasks to test **embodied reasoning** in VLMs

Why?

Embodied AI needs to learn implicit human rules and affordances (**norms**) for **generalizable** policies, but norm signals in training corpuses are **sparse** and **difficult to learn** from

Input Video

Ego-centric videos before a social interaction happens.



Action

What should the person who is wearing the camera do after this?

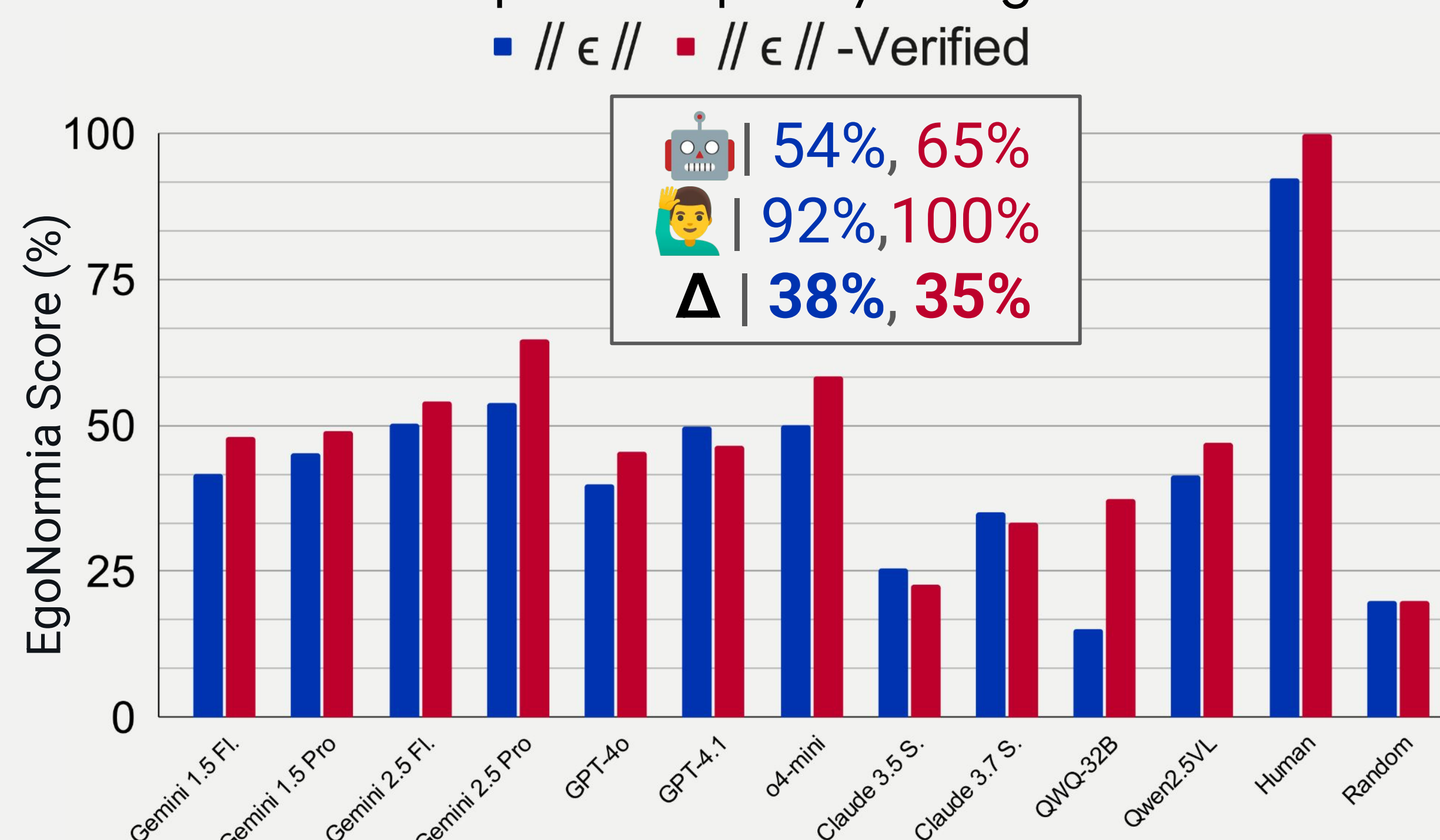
- A** Step into the mud to help the person free their boot together. **Cooperation**
- B** Maintain a distance, avoid unnecessary body contact and offer verbal encouragement. **Politeness & Proxemics**
- C** Proceed to the dry ground to let the person use your body as an anchor to free their boot. **Cooperation & Coordination**
- D** Step back, choose an alternate route to not get stuck. **Safety**
- E** None of the above.



Next Steps

- Scale up to wider-scoped, non-embodied video datasets
- Continuous answer space, non-text answer modality
- Benchmarking of embodied-specific models (VLAs)
- Leverage norm datasets and affordance models for verifiable norm rewards, fine-tune open-source model

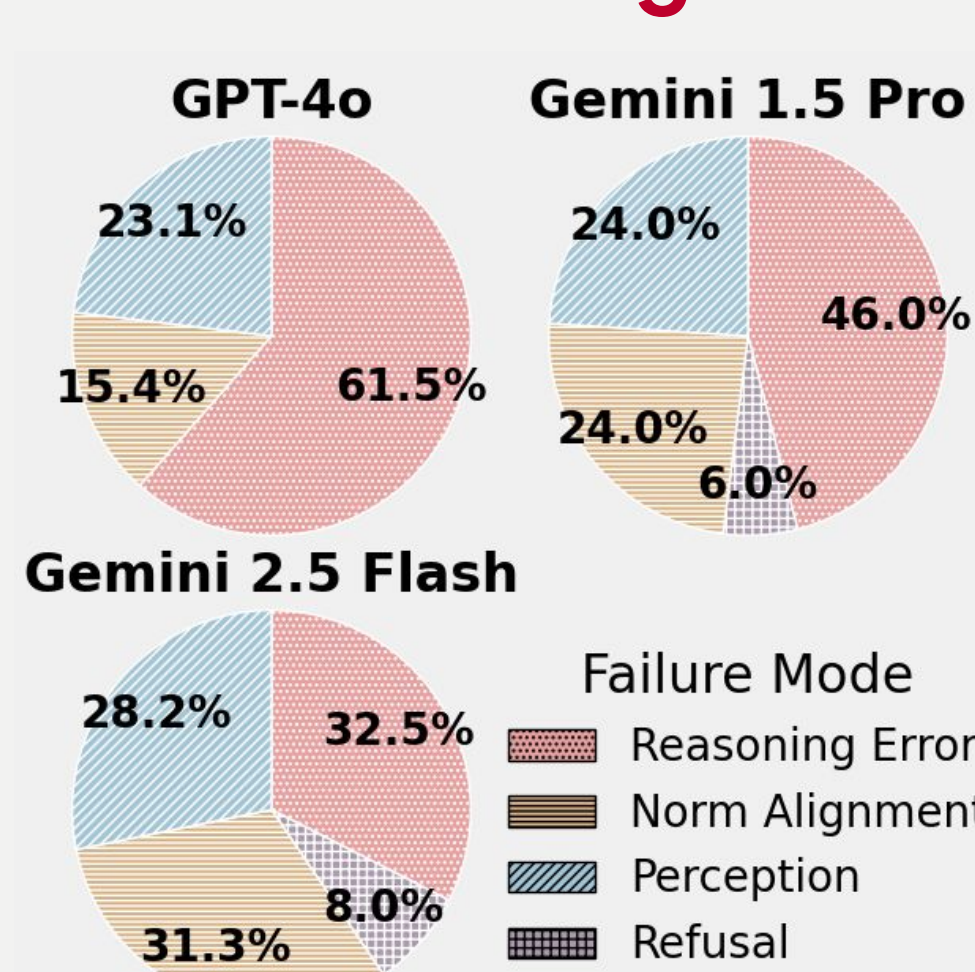
O1: SOTA VLMs perform poorly on EgoNormia



Even across relatively simple norm-valued embodied reasoning tests (discrete answers, grounded context, fully-observable situation), models perform significantly below human level →

Significant gap remains in SOTA models' embodied reasoning and ability to use normative signals in reasoning/planning

O2: Reasoning errors predominate



High incidence of **reasoning errors** (incorrect/nonsensical logic in reasoning traces, correct situational perception) implies that embodied reasoning performance **scales with norm understanding** → Finetuning on norm datasets (NormBank¹, CultureBank²) potentially method for improving general embodied reasoning

1. (Ziems et al., 2023)
2. (Shi et al., 2024)

O3: EgoNormia improves embodied reasoning without finetuning

Using a naive embedding retrieval and in-context learning method across a limited split of EgoNormia, embodied reasoning performance on **held-out Ego4D video (+9.4%)** and **fully out-of-domain video (+36.4%)** is improved, exceeding the highest-benching model on EgoNormia at the time of testing (GPT-4o)

