Scaling Synthetic Data with Soft Labels for Embodied Agent Alignment in CALVIN

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does scaling synthetic data with soft labels impact the alignment of embodied agents' actions with human preferences in the CALVIN validation set, measured by task success rate improvements. 15 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Designing for Human-Agent Alignment: Understanding what humans want from their agents. Research question: How does scaling synthetic data with soft labels impact the alignment of embodied agents' actions with human preferences in the CALVIN validation set, measured by task success rate improvements?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.2/10.

3 Results

12 papers retrieved. 15 claims extracted; 0 independently verified. Quality review score: 3.2/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The study consisted of 10 think-aloud interviews.	×	0.02
Participants reported being familiar with the concept of an agent performing tasks on the human's behalf.	×	0.11
The study was situated in the context of selling an item in an online marketplace.	×	0.03
Recruited participants reported having a less-than-satisfactory experience in an online marketplace negotiation while bu	×	0.02
The study received internal ethics approval from the organization.	×	0.02
Recruitment surveys asked about familiarity with the concept of an agent and past experiences transacting in an online m	×	0.03
A random sampling technique was used to select a subsection of survey respondents.	×	0.01
Quota sampling related to gender was used to shortlist participants.	×	0.02
11 individuals were shortlisted, including 1 backup participant.	×	0.02
Study sessions were conducted with 10 participants.	×	0.02
The study was conducted at a U.S.-based organization.	×	0.04
Participant roles included 1 UX designer, 1 prototyper, 3 researchers, 1 writer, 3 software engineers, and 1 data analys	×	0.03
Four of the 10 participants identified themselves as women.	×	0.03
One participant identified as African-American.	×	0.00
Four participants identified themselves to be of South-Asian descent.	×	0.03

References

- http://arxiv.org/abs/2506.16243v1
- http://arxiv.org/abs/2407.14477v4
- http://arxiv.org/abs/2404.04289v1