GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simón Posada Fishman, Marwan Aljubeh, Phoebe Thacker

Paper *Evals @ OAI*

Overview

GDPval is a **1,320 task benchmark** spanning **44 occupations** across the **top 9 U.S. GDP sectors**. It improves on prior evaluations by being more realistic, broader, and more long-horizon.

We published a **220-task open source gold subset**, drawn from the benchmark's broader task collection. The open-source release includes **task prompts and reference files for each task** in the gold subset. We also provide a **public automated grading service at** evals.openai.com to support broader benchmarking and reuse *(Figure 4)*.

Figure 1: GDPval sectors and occupations

Prof. services
- Software Developers
- Lawyers
- Accountants and Auditors
- Computer / Info Systems Managers
- Project Management Specialists

Healthcare
- Registered Nurses
- Nurse Practitioners
- Medical / Health Services Managers
- Supervisors of Admin. Support
- Medical Secretaries / Assistants

Retail trade
- Pharmacists
- General and Operations Managers
- Private Detectives / Investigators
- Supervisors of Retail Sales Workers

Finance
- Customer Service Reps
- Financial and Investment Analysts
- Financial Managers
- Personal Financial Advisors
- Financial Services Sales Agents

Wholesale trade
- Sales Managers
- Order Clerks
- Supervisors of Non-Retail Sales
- Sales Reps of Manufacturing, Technical / Scientific Products

Retail trade
- Mechanical Engineers
- Industrial Engineers
- Buyers and Purchasing Agents
- Shipping, Receiving, Inven. Clerks
- Production / Operations Supervisor

Government
- Recreation Workers
- Compliance Officers
- Supervisors of Police / Detectives
- Administrative Services Managers
- Social Workers

Real Estate
- Concierges
- Real Estate Sales Agents
- Real Estate Brokers
- Counter and Rental Clerks
- Community Association Managers

Information
- Producers and Directors
- Film and Video Editors
- Editors
- News Analysts and Journalists
- Audio and Video Technicians

Methodology

- **We prioritized economically important digital work** by selecting the top 9 by U.S. GDP contribution and then chose the highest-compensation predominantly digital occupations within each sector, yielding 44 occupations *(Figure 1)*.
- **We recruited experienced industry professionals** to author tasks, requiring at least 4 years of occupational experience plus additional screening; the average expert had 14 years of experience.
- **We built tasks from real work requests and deliverables**, often with reference files, and mapped them to O*NET occupational tasks to support broad and representative task coverage.
- **We applied an iterative quality-control pipeline** combining automated model-based screening with repeated human expert review across all 1,320 tasks in the full dataset *(Figure 2)*.
- **We evaluated the gold subset with blinded expert pairwise comparisons** and additionally trained an experimental automated grader to approximate expert judgments at lower cost *(Figure 3, Figure 4)*

Figure 2: Task creation / review process

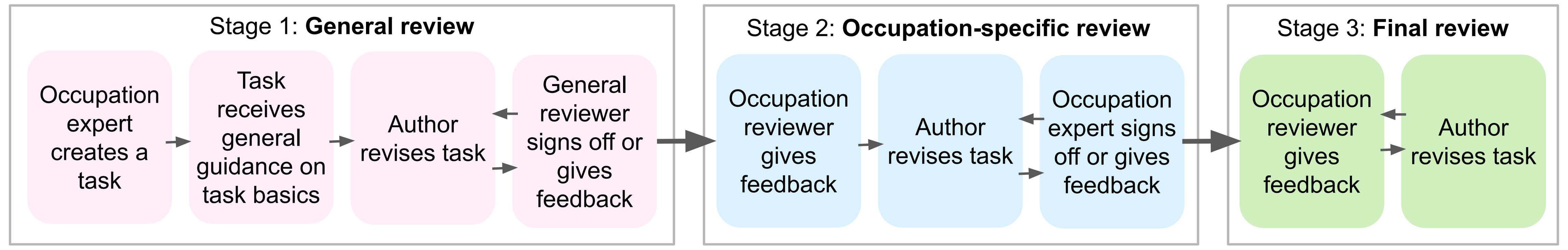


Figure 3: Pairwise Grading Setup

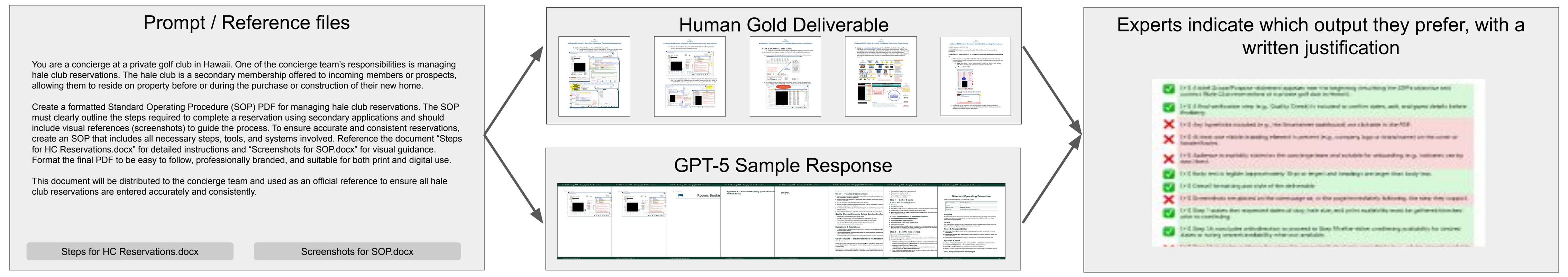
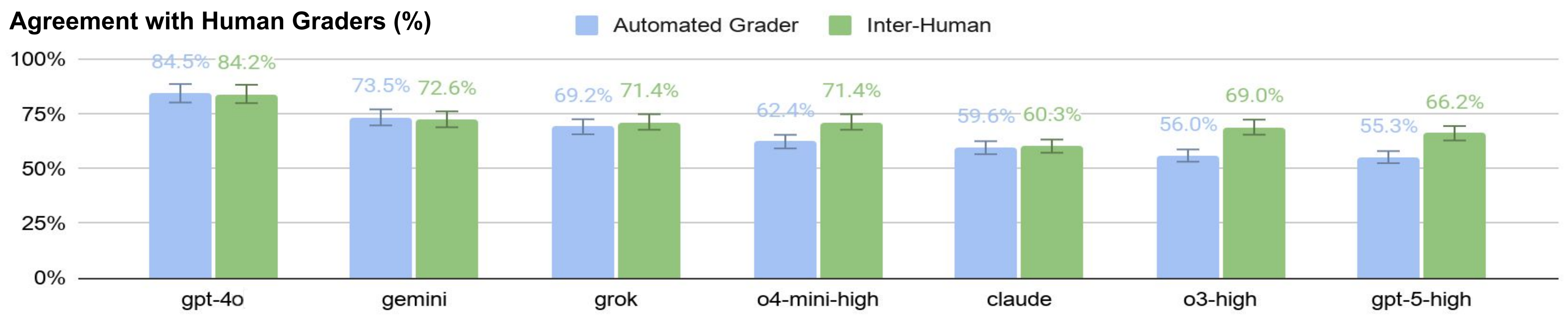


Figure 4: Automated Grader agreement with Humans



Findings

- **Updated frontier models now match or exceed expert-level deliverable quality in pairwise preferences,** with several models performing at above the 50% parity line in the updated results and the top model reaching roughly 74% wins + ties against industry experts *(Figure 5)*.
- **Models show different strengths and failure modes:** Claude performs especially well on visual and aesthetic quality, while GPT-5 is stronger on accuracy, instruction-following, and calculations with its losses more concentrated in formatting issues. Instruction-following remains the dominant failure mode across all models, with Gemini and Grok often missing requested deliverables or formats *(Figure 6)*.
- **Both oversight and scaffolding** improve practical performance: combining frontier models with expert review can reduce time and cost relative to unaided experts, while additional reasoning, prompting, and scaffolding further improve results, including a 5 percentage point gain in GPT-5 win rates with less problematic formatting artifacts *(Figure 7)*.

Figure 5: Frontier model performance

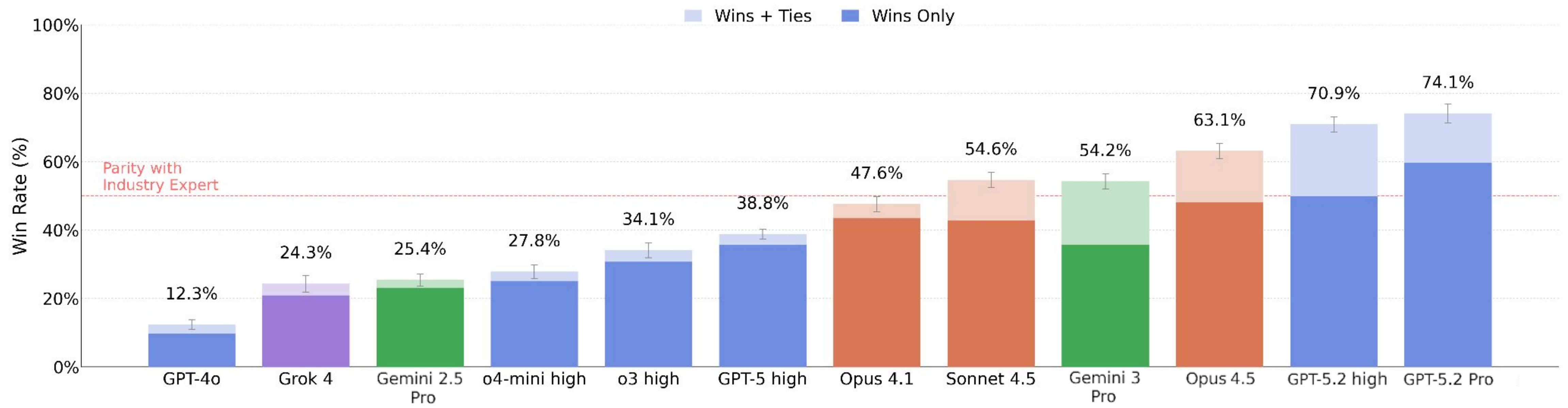


Figure 6: Error types by model

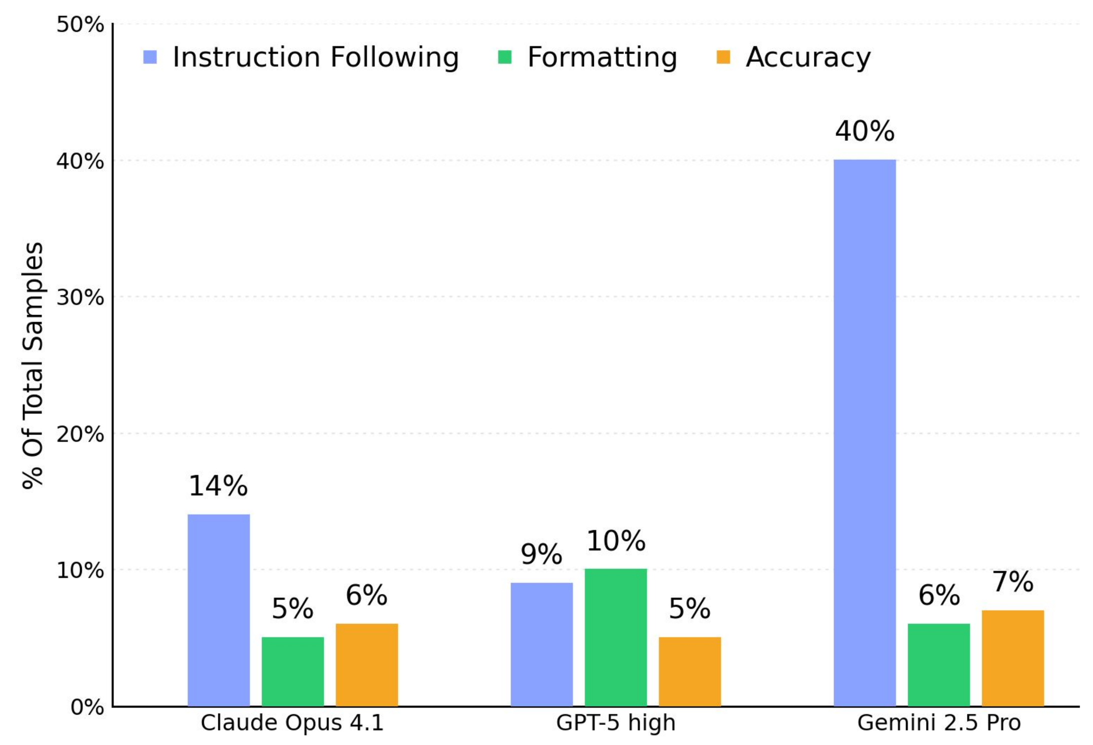


Figure 7: Speed / Cost comparisons