TACI.ai | TASK-AI CAPABILITY INDEX CONFIDENTIAL ADVISOR BRIEF | v0 .1 | JULY 2025

MEASURING LLM JOB-TASK PERFORMANCE: 180-Task Core Slice, 24-Hour Turn-Around

Author: Ely Baba, Founder, TACI.ai

Email: founder@taci.ai

Phone: +44 7387 167254

This document is confidential and intended solely for the recipient.

Abstract

TACI benchmarks frontier language models on a 180-task slice of real O*NET job statements that spans text, GUI, and vision. Each run finishes within twenty-four hours of a new checkpoint, with SHA-256-logged prompts, six-axis grading, and 95 % confidence intervals. The framework is adaptable: any occupation from the full ≈14 000-task O*NET taxonomy can be analysed by selecting the desired rows. This brief shows one occupation (Paralegals) and invites methodological feedback from researchers, model labs, and business decision-makers.

TABLE OF CONTENTS

1. WHY THIS MATTERS

Public model leaderboards still test synthetic trivia and academic exams. Businesses, AI labs, and policymakers need to know how new checkpoints perform on the *actual work tasks* that power the economy.

TACI answers that gap with a reproducible, twenty-four-hour pipeline over a 180-task core slice. Because every O*NET record is indexed, the system can be directed at **any occupation** in the U.S. Department of Labor taxonomy, today, by toggling the input manifest.

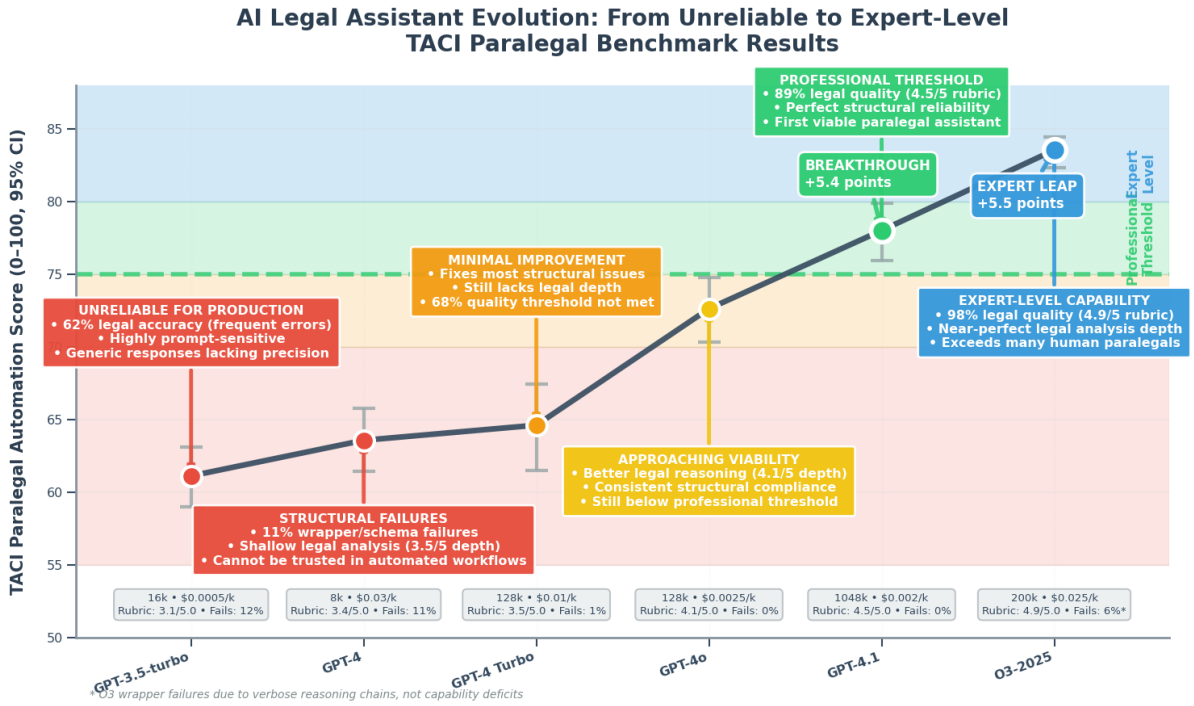
2. WHAT MAKES TACI DIFFERENT

- **Real-task grounding:** Importance-weighted O*NET tasks.
- **Multi-modal prompts:** Text, structured GUI actions, vision findings.
- **24-h refresh cadence:** New checkpoint, new numbers next day.
- **Transparent hashing:** Prompts, outputs, and grades SHA-256-logged.
- **Six-axis rubric + hard gates:** Quality *and* structural reliability.
- **Statistical discipline:** 1 000-bootstrap 95 % confidence intervals.

★ **Your input requested** A 30-minute call to critique design choices would be invaluable. If the approach proves useful, may we list you as an informal advisor (name & head-shot; no ongoing duties)?

3. SPOTLIGHT – PARALEGALS & LEGAL ASSISTANTS (SOC 23-2011)

Line chart of TACI paralegal scores for six GPT checkpoints; annotations mark structural failures, viability, expert level.



*Each point aggregates all 14 O*NET paralegal tasks with task-level importance weights. Results reflect 72 outputs per model (14 tasks × 2 temperatures × 3 prompt variants). Error bars show 95% confidence intervals computed from 1,000 bootstrap samples.*

• Reliability Cliff: The 11% Structural Failure Problem. GPT-4 has an 11% wrapper/schema failure rate, making it fundamentally unreliable for automated legal workflows, while GPT-4.1+ achieve perfect structural compliance. This is not just a performance gap; it's a trust threshold for real deployment.

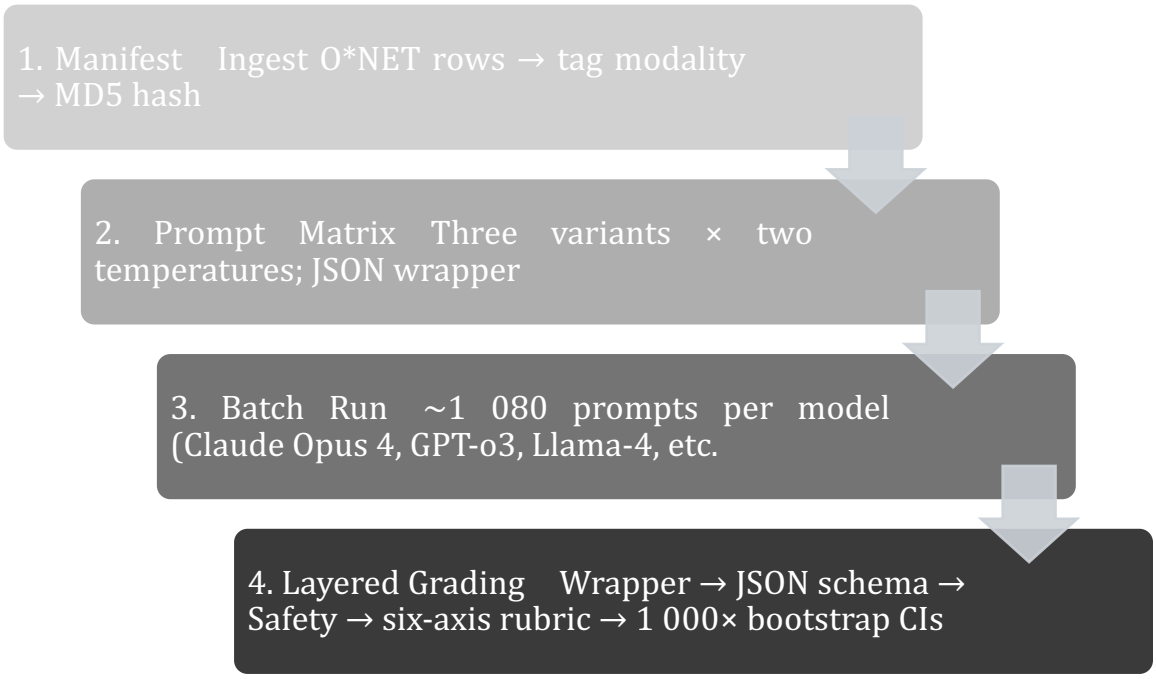
• Professional Viability Jump: GPT-4.1 as the Inflection Point. TACI pinpoints when AI becomes professionally viable: GPT-4.1 scores 78% with 4.5/5 rubric quality, crossing the "colleague-level" threshold. Below this, AI is only assistance-level; above, it can do independent paralegal work.

• Expert Paradox: O3's Tradeoff: Higher Quality, More Failures. O3 delivers 98% legal quality (4.9/5) and 83.5% score, despite 6% wrapper failures. GPT-4.1 has 0% failures

but lower quality. This shows reasoning depth and formatting compliance are separate axes: O3's verbose reasoning boosts analysis but breaks wrappers.

The identical analysis can be generated for *any* occupation or custom task list within twenty-four hours.

4. METHOD IN FOUR STEPS



Context-and-price bonus (v0 .1): Extra weight for models offering ≥ 128 k context at ≤ $0.002 / k tokens. (adjustable)

5. COMPARATIVE CONTEXT

Dimension	*HELM (Stanford '22)*	*MMLU (OpenAI '21)*	*BIG-Bench (Google '22)*	*TACI (2025)*
Real-world grounding	48 evaluation scenarios (HELM v2.1)	57 academic MCQs	399 synthetic tasks (BIG-Bench full)	**≈14 000 O*NET job tasks**
Modalities	Text	Text	Text	**Text + GUI + Vision**
Refresh cadence	Manual releases	Frozen snapshot	One-off release	**24-h automated core slice**
Longitudinal deltas	✗	✗	✗	**Built-in panel tracking**
Main use-cases	Ethics research	Education testing	Benchmark research	**Labour policy Business strategy Model eval**

HELM v2.1 (public release Oct 2024) contains 48 scenarios; the original 2022 paper listed 42. † BIG-Bench figure uses the full 399-task repository; many summaries cite the 204-task "Lite" subset.

6. CONTACT & NEXT STEPS

Ely Baba Founder, TACI.ai

founder@taci.ai

+44 7387 167254

To discuss or request additional occupations, please choose any open slot: https://calendly.com/founder-taci/30min