Domain-Specific Versus General-Purpose AI for European HTA Dossier Generation

A Human-Validated Benchmark Using Accepted G-BA Dossiers

Anton Wiehe · Pia Čuk · Susanne Schneller · Florian Woeste
Pharos Labs & SmartStep Consulting, Hamburg, Germany

HTA86
ISPOR 2026
Vienna, Austria
Nov 8–11

The Setup

European HTA submissions (German G-BA AMNOG & EU JCA) require synthesizing 1,000+ pages of clinical trial reports and SmPCs into structured Word templates.

Ground Truth: 8 authority-accepted G-BA AMNOG & CTD dossiers [1]
118,000+ words & 175 statutory tables

Massive Effort: Months of manual medical writing

Source Corpus: Hermetic trial CSRs, SmPCs & 85 CMC PDFs

Mandate for AI: Existing benchmarks test Q&A, not dossiers:

"Can domain-specific AI draft complete, citation-verified HTA dossiers matching accepted submissions?"

Methods

Tested Systems:
All systems received identical hermetic source PDFs with web search disabled:

Domain AI: Regulaido Engine (Slow & Fast) & Regulaido Agent
Coding CLIs: Claude Code CLI & OpenHands CLI
Tested LLMs: Gemini 4 Argon, Gemini 3.8 Flash, Opus 5.5 & Astra

Evaluation

Reference Audit (0–100): 80 pts section + 20 pts whole-doc coherence (anchored at 0.0 empty floor & 94.9 ceiling). [2]

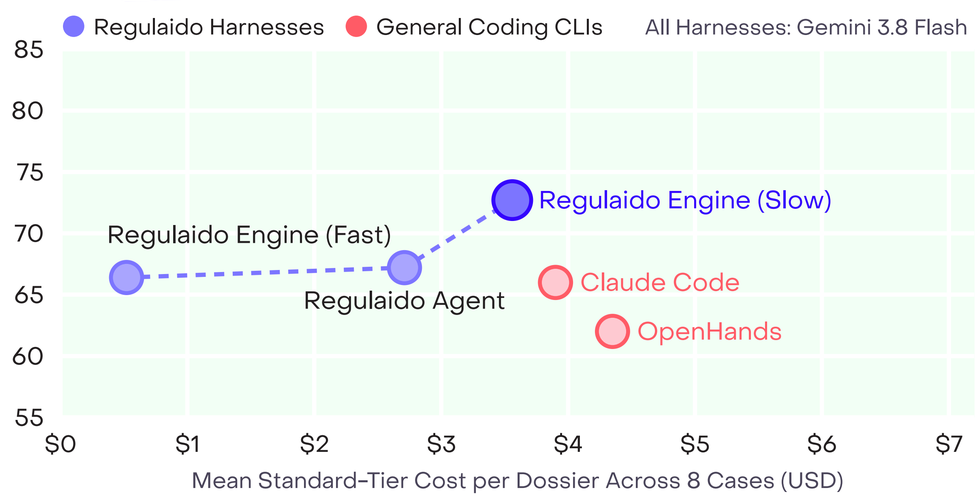
Human & Trap Calibration: 3-LLM jury calibrated via PPI to 180 ratings by 17 HTA experts + 47 planted error traps (97.9% recall). [2,3]

The Evidence

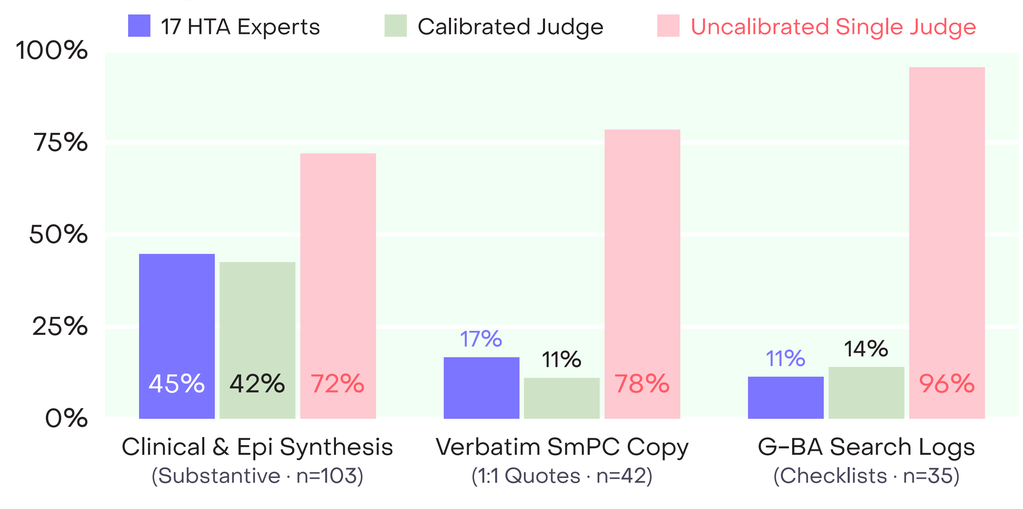
Agent Harness Comprison

Harness (Flash)	Score	Tables	Hallu.	Cost
Regulaido Engine (Slow)	72.7	100%	5.0	$3.56
Regulaido Agent	67.2	96%	7.0	$2.71
Regulaido Engine (Fast)	66.4	100%	6.8	$0.52
Claude Code	66.0	76%	8.5	$3.90
OpenHands	62.0	72%	11.8	$4.35

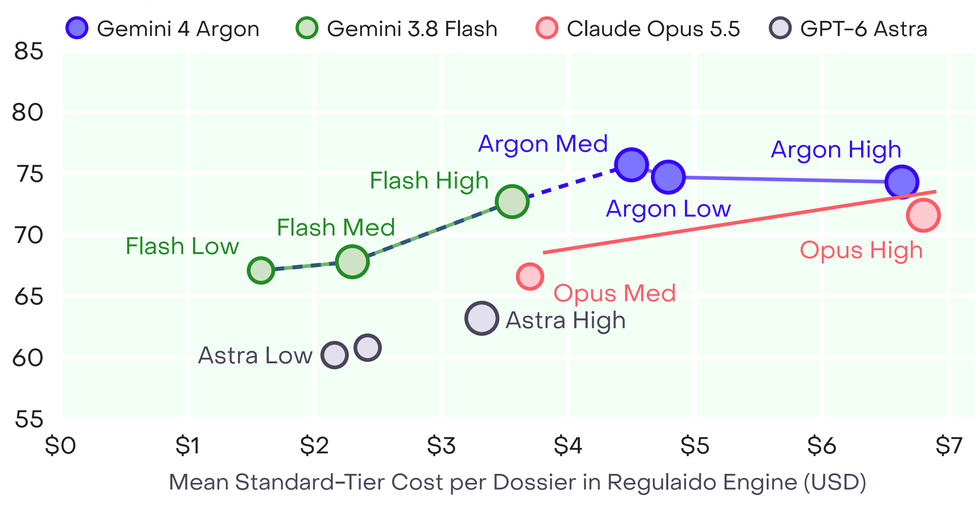
Agent Harness Quality (Pareto Frontier)



Human Expert vs. Judge Acceptance (N=180)



Score Across Frontier LLMs



The Impact

1. Domain Architecture Beats Coding CLIs by +6.7 to +10.7 Points at Lower Cost

Coding CLIs drop 24%–28% of statutory tables & figures (dumping numbers as un-auditable prose). Regulaido preserves 100% of tables with page-traced citations at $3.56 vs. $3.90–$4.35 (75.7 with Argon).

2. Gemini 3.8 Flash Delivers Outsized Cost-Effectiveness on HTA Synthesis

Despite ranking lower on general coding leaderboards, Gemini 3.8 Flash (72.7, $3.56) beats Claude Opus 5.5 High (71.6, $6.80) and GPT-6 Astra (63.2, $3.32) across all 8 dossiers.

3. Reference-Anchored Calibration Eliminates +84.1 pp Judge Bias

Single LLM judges accept 78%–96% of verbatim SmPC copy-paste vs. 11%–17% for 17 HTA experts;anchoring to accepted G-BA dossiers aligns grading within 2.2 pp.

4. Massive Time Savings: From Months to 35 Minutes ($3.56)

Synthesizes 1,000+ CSR & SmPC pages into 175 G-BA tables across 8 dossiers → HTA-Bench

Limitations

Engine Ceiling & Ongoing Optimization:
Regulaido Engine is still work in progress; ongoing section-agent and table synthesis upgrades unlock a higher quality ceiling.

Formatting, Figures & Citation Auditability:
The LLM jury observed only prose text and not pure tables with formatting and figures and do not check citations in detail.

Expanding Benchmark Tasks & Scope:
Adding more dossier tasks, drugs, EU JCA templates, and live registry lookups (RKI / Destatis) beyond these 8 hermetic cases.

Contact us
for full blog post & live leaderboard

References:
1. G-BA Verfahrensordnung Modul 1-4 (2026).
2. Angelopoulos et al. Science 382:669 to 674 (2023)
3. Verga et al. arXiv:2404.18796 (2024).

anton.wiehe@pharos labs.com
benchmarks.regulaido.ai