Per-script evaluation accounting

Companion to the Methods (How it was judged) and the Results of the manuscript. One row per scripted consultation; classifications are the harness's own (scripts/evaluate_urgency.py); outcomes are read from the committed evaluation records (evals/), not re-run. All results are script text fed to the engine at deterministic settings (temperature 0, seed 42), one run per script.

Table S1. Per-script evaluation accounting.

Script	Type (harness)	Harness(es)	Urgency outcome	Note outcome
01_chest_pain_en	Emergency	urgency, notes	Fired turn 4/40, cleared — pass	29/29 cited
02_febrile_child_en	Routine	urgency, notes	False alarm (fired turn 8/33, cleared) — KEPT failure	29/29 cited
03_diabetes_review_en	Routine	urgency, notes	Silent — pass	34/34 cited
04_asthma_en	Routine	urgency, notes	Silent — pass	30/30 cited
05_epigastric_pain_en	Routine	urgency, notes	Silent — pass	37/37 cited
06_tia_funny_turn_en	Emergency	urgency, notes	Fired turn 4/39 — pass	30/30 cited
07_gi_bleed_en	Emergency	urgency, notes	Fired turn 8/31 — pass	23/23 cited
08_pulmonary_embolism_en	Emergency	urgency, notes	Fired turn 4/32 — pass	36/36 cited
09_septic_child_en	Emergency	urgency, notes	Fired turn 4/27 — pass	30/30 cited
10_testicular_torsion_en	Emergency (held-out; authored after the harness was frozen; n=1)	urgency addendum, notes	Fired turn 4/38 — pass	23/24 cited; the uncited claim tagged "(uncited — verify manually)" and gated at approval
11_tired_all_the_time_uk	Routine (negative control)	restraint	Silent — pass	Restraint arm: KEPT failure (unearned

				confidence)
12_dizziness_uk	Routine (negative control)	restraint	Silent — pass	Restraint: pass
13_migraine_uk	Routine (differentiable)	restraint	Silent — pass	Restraint: pass
14_giant_cell_arteritis_uk	Emergency (buried)	restraint	Fired turn 4/31, cleared by 28 — pass	Restraint: pass
15_cauda_equina_uk	Emergency (buried)	restraint	Fired turn 8/27, cleared by 24 — pass	Restraint: pass

Headline arithmetic, reconciled. Emergencies: 8 of 8 fired and cleared (01, 06, 07, 08, 09, held-out 10, buried 14, 15). Routine: 6 of 7 silent (03, 04, 05, 11, 12, 13), with the one false alarm (02) recorded and kept. Note citation: 301 of 302 claims across the ten note-generation scripts resolve to valid transcript turns; the single exception is shown being caught by the machinery.

Room-run provenance (Table 2 of the manuscript). Severity is the clinician's judgement of clinical consequence; no formal severity scheme was applied. Run-level system versions: the five runs executed on the 2026-07-28 build of the supervised history-taking phase, per the engineering record.