A Reproducible, Provenance-Preserving Architecture for Sequential Intervention-Threshold Evaluation in Large Language Models

Alex Kolesnikov
Eton College, Windsor, Berkshire, United Kingdom
E-mail: alexkolesnikov2010@gmail.com

Keywords: large language models, sequential evaluation, first-event analysis, event-history data, reproducibility, provenance

Received: August 25, 2026

Sequential evaluation of large language models (LLMs) needs an auditable way to analyse when intervention first occurs. The architecture presented here uses cumulative scenarios and reconstructs first-ACT/censoring and at-risk datasets from a master record for discrete-time event-history analysis. Its methodological contribution is the integration of established first-event reconstruction with a frozen instrument, independent current-state calls and serving provenance, while retaining later responses for audit. In a thirteen-model demonstration, 26,000 stage-level records yielded 25,999 decision-valid responses, 3,120 scenario-runs and 10,088 at-risk rows. The derived endpoint files and all 2,000 prompt coordinates were regenerated exactly from deposited records. Complete conditional hazards were stage-dependent but non-monotone; within-group linear trend estimates were OR 12.591 (95% CI 8.481–18.694) in Group A, 10.346 (9.292–11.518) in Group B and 4.377 (3.815–5.022) in Group C, and a quadratic sensitivity confirmed nonlinearity. Later WAIT responses occurred in 21.05% of ACT-containing runs. The reconstruction checks establish analytical reproducibility of one realised first-intervention path under recorded serving conditions; they do not establish a stable latent threshold, normative safety validity or exact recreation of historical provider responses.

Povzetek: Članek predstavlja metodo za preverljivo in ponovljivo analizo, kdaj se veliki jezikovni modeli v postopno razvijajočih se scenarijih prvič odločijo za ukrepanje.

1 Introduction

Large language models are increasingly used as advisory systems in settings where information arrives progressively rather than as a completed case. In health triage, cybersecurity monitoring, fraud review and emergency communication, the operational question is not only whether a final state appears risky, but when the available evidence becomes sufficient to justify action. Static benchmarks generally reveal the complete item before scoring one answer and therefore cannot distinguish a system that would act at an early warning sign from one that waits until danger is almost explicit (Liang et al., 2023; Srivastava et al., 2023). For sequential advisory systems, first-intervention timing is itself a measurable behavioural property.

This timing problem is also an evaluation-design and data-structuring problem. Sequential decision methods treat action as a stopping event: ordered evidence states are observed until a decision criterion is crossed, while runs without an event inside the observation window remain right-censored (Wald, 1945; Green and Swets, 1966; Allison, 1982; Singer and Willett, 2003). Applying that logic to served LLMs requires more than assigning a stage number. Cumulative evidence must be presented without allowing an earlier model decision or explanation to become new evidence; the authoritative source record must remain distinguishable from the derived event-history risk set; and mutable provider routes, endpoint aliases, parser outcomes and model identities must remain auditable.

Sequential and stateful LLM evaluation is established prior work. HELM standardises broad multi-metric evaluation, MT-Eval and MT-Bench-101 assess persistent multi-turn dialogue capability, and AgentBench, tau-bench and ToolSandbox evaluate interactive or stateful agent behaviour (Liang et al., 2023; Kwan et al., 2024; Bai et al., 2024; Liu et al., 2024; Yao et al., 2025; Lu et al., 2025). These approaches principally score dialogue capability, task completion, final environment state, repeated-trial reliability or trajectory milestones rather than a right-censored first-intervention endpoint under a frozen escalating evidence sequence. Prompt-sensitivity and reporting work further show that wording, implementation and provenance choices can materially affect interpretation (Sclar et al.,

2024; Gallifant et al., 2025; Aloqalaa, Soiland-Reyes and Goble, 2026; Edmunds et al., 2026). Table 8 in Section 4.2 compares these approaches in terms of their evaluation targets, first-event handling, provenance and analytical reproducibility.

The first-event reconstruction used here also has prior methodological lineage. Stopping rules, right-censoring and discrete-time at-risk representations are established methods, and Kolesnikov (2026b) previously published a reproducible pipeline that converts human stopping-rule survey exports into first-action/censoring records and stage-level at-risk rows. The present paper treats that wide-to-long stopping-rule transform as prior work. Its scope is the extension of that lineage to mutable served LLM systems, where independent cumulative calls, exact prompt/parser reconstruction, provider-route/model-identity provenance and deliberately retained post-first-ACT model outputs create additional evaluation requirements. The provenance-preserving master record and deterministic first-event derivatives are inherited and adapted components; the served-system controls and post-event model audit are the system-specific extensions.

1.1. Research question, goals and evaluation criteria

The study therefore asks how sequential LLM decisions can be represented so that first-intervention timing is analytically valid, deterministically reconstructable and auditable across mutable serving routes. The architecture is assessed against three linked goals: (1) factorial completeness and exact post-collection endpoint reconstruction; (2) prompt, parser, route and model provenance sufficient to trace every analytical record to its source coordinate; and (3) a separate descriptive check that the frozen scenario instrument elicits stage-dependent ACT behaviour. A thirteen-model cohort provides the demonstration rather than a permanent leaderboard.

The first two goals are evaluated through completeness and exact-match checks, source-to-derivative links, parser audits and recorded route/model identities (Sections 2.6–2.9 and 3.1; Tables 4–5). Required invariants include one endpoint record per run, one event per non-censored run, and no at-risk row after first ACT. For the third goal, stage-dependent ACT is an instrument-response expectation examined through complete conditional hazards and within-group trend and shape analyses (Section 3.3; Table 7), without assuming monotonic hazards or a normatively correct intervention stage. The revised primary and quadratic specifications were introduced during revision and are not presented as pre-specified confirmatory tests.

2 Materials and methods

2.1. Terminology and analytical unit

A cumulative scenario state was a current prompt in which previously introduced warning signs remained active unless explicitly resolved and later stages added, intensified or maintained evidence. A stateless current-state call was a fresh API request containing that cumulative state but not the model's earlier WAIT/ACT outputs, confidence scores or explanations. A scenario-run was the ordered set of stage calls sharing model, scenario, prompt role, response mode and sampling condition. ACT was an exact parsed decision meaning intervene, escalate or trigger the domain-appropriate response now; WAIT meant continue monitoring. A response was decision-valid when the fixed parser recovered an exact WAIT or ACT value, while strict compliance additionally required the whole output to match the requested schema.

First-ACT was the earliest decision-valid stage within a scenario-run at which ACT appeared. A run with no decision-valid ACT by the final displayed stage was right-censored. An at-risk row was a decision-valid stage observed before the run had already experienced first ACT, including the first-ACT stage itself when an event occurred. In this empirical cohort, first-ACT is interpreted as one realised response path under the recorded route, collection window, prompt condition and provider-standard sampling condition; it is not assumed to be a stable latent threshold of the underlying model or a normatively optimal intervention point.

2.2. Evaluation objects, software and release structure

Analyses were conducted in R 4.6.0; cluster-robust covariance used sandwich 3.1.1, and the final Figure 4 renderer uses Python/matplotlib as documented in the public package. Here, analytical reproducibility means deterministic regeneration of derived datasets, prompt coordinates, release checks, tables and figures from deposited post-collection records. It does not imply exact recreation of historical provider responses: collection of a new cohort still requires provider-specific execution, and endpoints, wrappers and serving infrastructure can change. The package contains a one-command runner (Rscript 00_run_all.R) that reconstructs endpoints, refits the reported models, regenerates every manuscript table and Figures 1–4, builds a metric registry and validates generated outputs against saved references. Table 1 summarises the principal released objects.

Table 1: Required resources and principal data-architecture outputs, with their roles and locations in the reproducibility package.

Resource	Role in the data architecture	Released location
Scenario instrument and prompt manifests	Frozen staged scenarios, hidden construction coordinates, prompt wording and template checks.	02_instrument_and_prompts/scenario_instrument_manifest.csv; frozen_prompt_manifest_2000_rows.csv; prompt_templates/
Master stage-level dataset and codebook	Authoritative source record containing every canonical administered stage and output plus route-specific audit fields.	01_data/d2_master_stage_level_13_models.csv; 01_data/CODEBOOK.md
Scenario-level dataset	One realised first-ACT or right-censoring outcome per scenario-run.	01_data/d2_first_act_scenario_level_13_models.csv
At-risk dataset	Rows retained through first ACT, with one event per non-censored run.	01_data/d2_stage_level_at_risk_hazard_dataset.csv
Freeze, inclusion and audit records	Canonical source mapping, model inclusion/exclusion, parser exceptions and identity recovery.	01_data/d2_model_inclusion_table.csv; d2_excluded_deferred_models.csv; d2_canonical_source_file_manifest.csv; d2_parse_compliance_exceptions.csv; d2_openai_returned_model_recovery.csv
One-command reconstruction and outputs	Reconstruct endpoints, refit models, regenerate tables/Figures 1–4 and validate saved outputs.	00_run_all.R; 03_code/; 05_outputs/; 07_documentation/

The demonstration contains 26,000 frozen stage-level coordinates and canonical response records. This count is not identical to total provider attempts because some coordinates required transport recovery or additional execution attempts. Released usage fields contain 10,215,059 input tokens and 2,425,533 output tokens. Provider-reported total-token fields sum to 14,997,588, but routes differ in whether reasoning, cache or other categories are included, so no false common cost or sampling parameter is reconstructed. The locally retained canonical raw-response archive contains 26,039 JSON files, including retry and transport artefacts associated with included runs.

The release is organised around 01_data/d2_master_stage_level_13_models.csv as the source of record. The scenario-level and at-risk files are deterministic products of this master record rather than independently editable datasets. 02_instrument_and_prompts/frozen_prompt_manifest_2000_rows.csv supplies one exact prompt for every unique scenario-stage-role-mode coordinate, including the 6,000 OpenAI rows whose merged prompt_text field is blank. Public endpoint reconstruction, prompt checks and manuscript outputs require no provider access; raw provider responses retained locally are not represented as publicly reproducible evidence unless specifically deposited.

2.3. Study workflow

The six-phase workflow is shown in Figure 1, and its inputs, operations and outputs are summarised in Table 2.

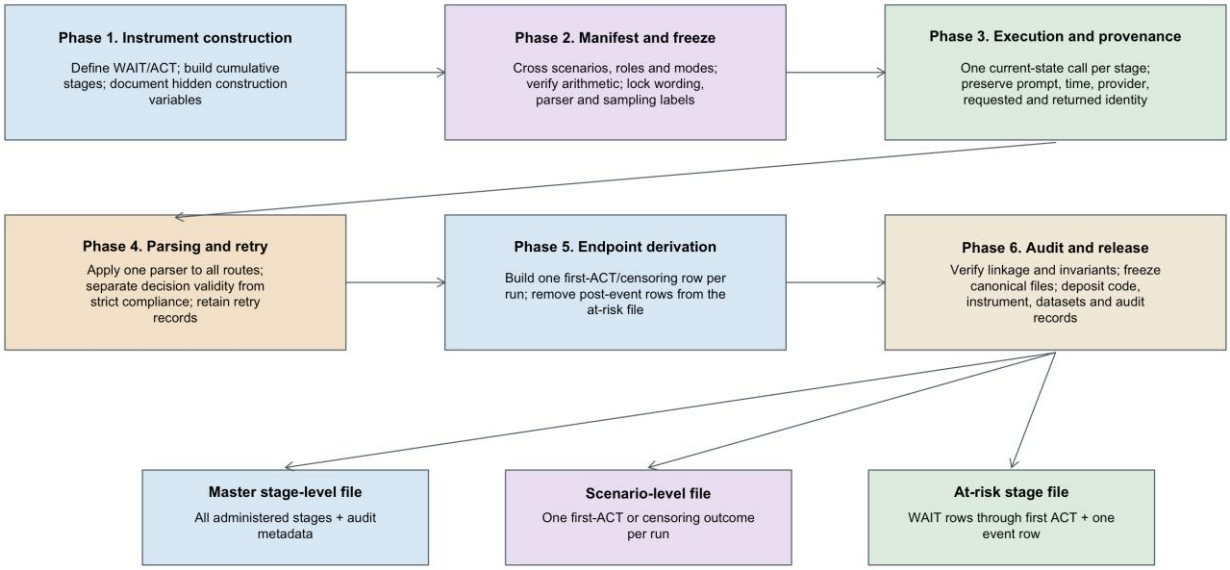

Figure 1: Six-phase study workflow and the three linked data layers.

Table 2: Inputs, operations and outputs of the six-phase workflow

Phase	Input	Operation	Output
1. Instrument construction	Decision definition and domain materials	Build cumulative staged states; document hidden construction variables and matched trajectories.	Versioned scenario instrument
2. Manifest and freeze	Instrument, prompt roles and response schemas	Generate all factor combinations; verify counts; lock wording, parser and sampling labels.	Immutable coordinate manifest
3. Execution and provenance	Frozen manifest and endpoint access	Issue one current-state call per coordinate; preserve prompt, time, route and identity metadata.	Canonical stage responses
4. Parsing and audit	Model outputs and fixed parser	Apply one parser; separate decision validity from strict compliance; preserve retry/exception records.	Audited master stage-level file
5. Endpoint derivation	Valid ordered master rows	Locate first ACT or censoring; retain only stages at risk through the event.	Scenario-level and at-risk files
6. Reproduction and release	Canonical sources and derived outputs	Verify invariants/linkage; regenerate analyses, tables and figures; deposit documentation.	Analytically reproducible release

2.4. Instrument construction

2.4.1. Decision task and unit of observation

WAIT and ACT were fixed before scenario construction. The master-file unit was one served-model response to one scenario stage under one prompt-role and response-mode condition; the scenario-run unit was the ordered set of stages sharing model, scenario, prompt role, response mode and sampling condition. ACT was operational rather than normative: it represented a recommendation to intervene, escalate or trigger the relevant response now, not evidence that the intervention was objectively correct.

2.4.2. Cumulative states and matched trajectories

Scenario stages were constructed so that unresolved signs persisted across stages, with later stages adding, intensifying or maintaining evidence. The instrument contains 24 scenario instances arranged as twelve gradual/sudden matched pairs. Group A reuses four five-stage health pairs from the published human sequential-decision instrument (Kolesnikov, 2026a). Group B contains four ten-stage synthetic health pairs covering respiratory illness, abdominal illness/dehydration, head injury/neurological symptoms and allergic reaction/breathing difficulty. Group C contains four ten-stage non-health pairs covering cybersecurity breach, financial fraud, domestic natural-gas leak warning and weather/flood warning. Each pair has gradual and sudden trajectories, giving eight scenarios per group. Table 3 summarises the scenario-state and prompt-policy construction rules.

Table 3: Scenario-state and prompt-policy construction rules, with implementation details and constraints for reproducing the frozen instrument.

Design component	Demonstration implementation	Reproducibility constraint
Scenario groups	A: eight five-stage health scenarios; B: eight ten-stage synthetic health scenarios; C: eight ten-stage non-health scenarios.	Version group structure; do not treat unequal stage windows as a common severity scale.
Cumulative persistence	Current prompt states that earlier signs continue and adds/intensifies evidence.	Never imply resolution unless explicitly stated; fully enumerated persistence is a separate version.
Hidden construction logic	Signal IDs, weights, cumulative scores, bands, pair IDs and trajectory labels.	Document for audit but do not expose to the model.
Prompt policy	Five fixed roles crossed with direct and deliberative response schemas.	Treat each wording as a specific condition; paraphrases/translations require a new version.
Stopping decision	Exact WAIT or ACT plus confidence; deliberative mode adds two visible considerations.	Fix schema before collection; do not interpret visible considerations as private chain-of-thought.

A hidden score was used only to regularise scenario construction and was not presented as a validated operational scale. Signal weights and internal bands were synthetic construction parameters, not clinical NEWS2 values; aggregate early-warning-score logic provided only a conceptual inspiration for combining signals (Royal College of Physicians, 2017). Models saw only the scenario text, stage number, role instruction and response schema; signal IDs, weights, cumulative scores and severity bands remained hidden construction metadata.

Five fixed prompt roles modified the requested decision policy while leaving scenario evidence unchanged: neutral, safety-oriented, resource-constrained, expert-risk-assessor and evidence-conservative. They were chosen to operationalise distinct decision emphases rather than universal psychological constructs. Each wording is therefore interpreted as a specific experimental condition; paraphrases or translated versions would require a new frozen instrument.

Response format was also frozen. Direct mode requested only decision and confidence fields. Deliberative mode requested two brief decision-relevant considerations before the same decision and confidence fields; these visible considerations were treated as ordinary generated text, not as private chain-of-thought. Both modes used constrained JSON, and the same parser rules were applied across providers.

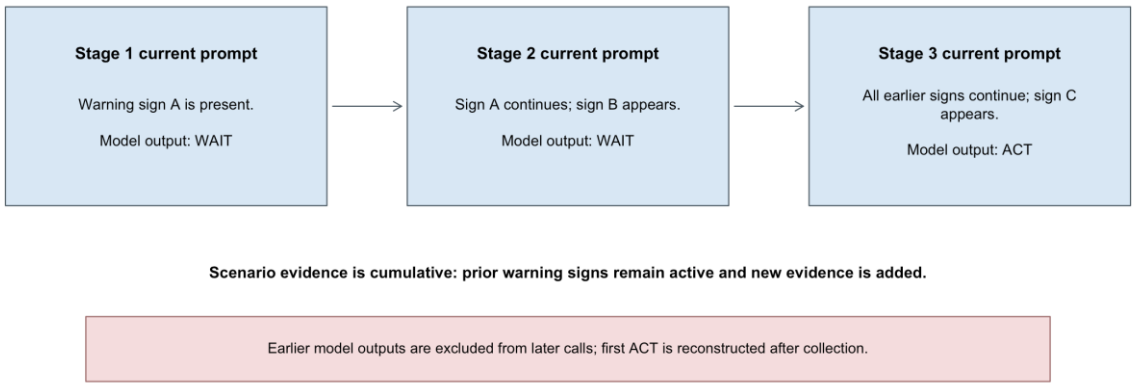

Figure 2: Stateless presentation of cumulative scenario states. Previous model outputs are excluded from later inputs by construction while current evidence remains cumulative.

2.4.3. Stateless presentation of cumulative evidence

As illustrated in Figure 2, each stage was submitted as an independent current-state call. This prevents an earlier decision, confidence score or explanation from entering the next input by construction; the study does not infer the counterfactual behavioural effect of persistent conversational history. Cumulative persistence was sometimes represented compactly (for example, "all earlier signs continue") rather than by verbatim repetition. That convention was frozen consistently to keep each call self-contained, but it relies on the persistence instruction being interpreted as intended; a fully enumerated cumulative version is therefore a separate robustness condition rather than a silent replacement of the frozen instrument.

2.5 Manifest generation, calibration and freeze

Calibration occurred before the primary cohort. A diagnostic model excluded from the final comparison was used to detect ambiguous wording, parser failures and severe floor/ceiling behaviour. Two pre-freeze changes were documented: plateau wording was strengthened after the diagnostic model treated unresolved signs as improving, and an ambiguous binary-output instruction was clarified. No included-model results were used to tune scenario wording.

After the freeze, scenarios, role wording, output schemas, parser criteria, sampling labels and first-event definitions were not altered in response to included models. Unexpectedly early action, late action, refusal or formatting failure was treated as an outcome or audit event rather than a reason to rewrite the instrument. This freeze separates instrument construction from model comparison.

The complete row-level manifest was generated before API execution. Every row specified scenario group, pair ID, scenario ID, domain, trajectory, stage, role, response mode, stage text and full prompt, with hidden construction fields retained only in non-model-facing columns. Crossing 200 ordered scenario-stage states with five roles and two response modes yielded exactly 2,000 prompt coordinates per model.

The frozen arithmetic provided a transparent completeness check. Eight five-stage Group A scenarios crossed with five roles and two modes yielded 400 rows per model; eight ten-stage Group B scenarios and eight ten-stage Group C scenarios each yielded 800 rows, for 2,000 rows per model and 26,000 canonical records across thirteen models.

Identifiers were separated by function. scenario_id identified the scenario version; pair_id linked gradual and sudden versions; the compound run key linked all stages for one model-condition run; api_call_id identified one frozen coordinate; and analysis_model_label identified the canonical served-model run used in analysis.

2.6 Time-stamped model execution and provenance

Execution iterated over the frozen manifest with one model-facing call per coordinate. The collection code preserved the exact prompt and, where exposed, provider, requested identifier, returned identifier, start/finish time, response identifier, output text, token fields, route-specific sampling metadata and retry flags. All 26,000 canonical rows carry the recorded standard sampling condition; a common numeric temperature and common provider seed are not available in the merged release and are not inferred retrospectively.

The cohort was collected from 25 June to 7 July 2026 and contains thirteen canonical completed runs across

OpenAI, Anthropic, Google, xAI, Mistral and Alibaba/DashScope routes (Table 4). OpenAI contributes public row-level start/finish timestamps for 6,000 rows; for other routes, the merged public file supports only the overall cohort collection window. Attempted, incomplete, region-restricted, quota-limited and superseded runs are recorded separately rather than pooled into the canonical cohort.

Raw provider responses were retained locally where feasible because they support parser audits and identity recovery. OpenAI returned identities were author-verified from retained raw response bodies and the derived recovery record is public, but the underlying raw bodies are not generally deposited. Returned-model fields for other routes are visible in the released canonical record, although independent raw-response verification is not uniform. The package therefore distinguishes publicly reconstructable analytical provenance from author-retained provider evidence.

Provider failures and model exclusion decisions remained in the audit trail. Gemini 3.1 Pro Preview was incomplete because of demand/quota constraints, Sakana Fugu Ultra was region-restricted, and superseded Sonnet, Opus and GLM runs were excluded from canonical datasets. No incomplete or superseded run was silently pooled into the final cohort.

Serving metadata were interpreted conservatively. A marketing label can resolve to an endpoint alias, and a hosted route can add wrappers or safety layers that differ from direct access. Provider, requested identifier, returned identifier where available, response ID and collection timing are therefore separate provenance fields. The unit of interpretation is the served system observed through a particular route and time window, not an assumed immutable checkpoint.

2.7 Deterministic parsing and predefined retry handling

The same deterministic parser was applied to all models. The entire response was first parsed as JSON; if that failed, one documented recovery routine extracted a single candidate JSON object. Only exact WAIT or ACT values were decision-valid. Confidence had to be numeric and between 0 and 100, and deliberative outputs had to contain at least two non-empty considerations. Free text was never manually converted into a decision.

Strict compliance was separated from decision validity. Strict compliance required the whole response to satisfy the requested schema, whereas decision validity required a recoverable valid WAIT/ACT value under the documented parser. A decision-valid non-strict row could therefore enter endpoint construction while remaining identifiable for sensitivity analysis. Transport recovery and semantic re-prompting were conceptually distinct, but the merged attempts_used field is route-specific, blank for 8,000 rows and records total attempts rather than retry type; it is not interpreted as repeated stochastic sampling.

The fixed parsing policy prevents two opposite sources of discretion: excluding every harmless wrapper can create model-specific missingness, while manually repairing arbitrary prose can inject investigator judgement. The deposited parser audit records whole-response parsing, extracted-object recovery, field validity and strict-schema compliance. Invalid responses remain in the master record rather than being hand-coded.

Implementation safeguards were pre-specified. A substantive provider failure could be retried without changing the frozen coordinate, while a semantic clarification was permitted only under the documented refusal rule; a 20% model-by-domain refusal threshold would have triggered one clarification and possible cell exclusion, but the safeguard was not triggered. Duplicate coordinates or partial canonical runs were treated as release failures rather than silently reconciled. Table 4 reports the canonical model cohort against which these rules were applied.

Table 4: Canonical thirteen-model cohort and serving provenance

Analysis label	Provider / route	Requested -> returned model	Timing	Sampling
alibaba_hosted_deepseek_v4_pro_us	Alibaba/DashScope-hosted	deepseek-v4-pro-us -> same	Cohort window	standard
alibaba_hosted_glm_5_2_rerun3	Alibaba/DashScope-hosted	glm-5.2 -> same	Cohort window	standard
alibaba_hosted_kimi_k2_5	Alibaba/DashScope-hosted	kimi-k2.5 -> same	Cohort window	standard
alibaba_qwen_3_7_max_2026_06_08	Alibaba/DashScope	qwen3.7-max-2026-06-08 -> same	Cohort window	standard
anthropic_claude_fable_5	Anthropic	claude-fable-5 -> same	Cohort window	standard
anthropic_claude_opus_4_8_rerun2	Anthropic	claude-opus-4-8 -> same	Cohort window	standard
anthropic_claude_sonnet_4_6_rerun2	Anthropic	claude-sonnet-4-6 -> same	Cohort window	standard
google_gemini_2_5_pro	Google	models/gemini-2.5-pro -> gemini-2.5-pro	Cohort window	standard
mistral_large_3_2512	Mistral	mistral-large-2512 -> same	Cohort window	standard
openai_gpt_5_4	OpenAI	gpt-5.4-2026-03-05 -> same†	29 Jun 12:41-14:16 BST	standard
openai_gpt_5_4_mini	OpenAI	gpt-5.4-mini-2026-03-17 -> same†	29 Jun 10:37-12:20 BST	standard
openai_gpt_5_5	OpenAI	gpt-5.5-2026-04-23 -> same†	29 Jun 15:06-17:21 BST	standard
xai_grok_4_3	xAI	grok-4.3 -> same	Cohort window	standard

Analysis labels reproduce the canonical identifiers used in the deposited analytical records. "Same" means that the returned model identifier matches the requested identifier. "Cohort window" denotes 25 June–7 July 2026, where exact row-level timing is unavailable in the merged public file. OpenAI collection intervals are shown for 29 June 2026; exact timestamps remain available in the deposited records. All runs use the recorded provider-standard sampling condition; no common numeric temperature or seed is inferred.

† OpenAI returned identities were recovered and author-verified from retained raw response bodies. The derived recovery record is publicly deposited; the underlying raw bodies are not generally deposited.

2.8. First-event endpoint construction

Within each scenario-run, decision-valid rows were ordered by stage. Let D_rs denote the parsed decision for run r at stage s. The first-event time was T_r = min{s: D_rs = ACT}. If no decision-valid ACT occurred by the final displayed stage S_r, the run was right-censored and first_act_stage remained missing. A missing first-ACT stage therefore represents censoring rather than ordinary missing data.

For a non-censored run, stages 1 through T_r were retained in the at-risk file and event = 1 was coded only at T_r; earlier WAIT rows received event = 0. For a censored run, all decision-valid observed stages through S_r were retained with event = 0 throughout. Post-first-ACT rows remained in the master file for audit but were excluded from the at-risk file because the first event had already occurred. This exclusion follows from the definition of the first-event estimand, irrespective of whether later independent judgements reverse.

The scenario-level file retains the coordinates required to join back to the master record: model label, scenario group, pair and scenario IDs, domain, trajectory, prompt role, response mode, sampling condition, first-ACT stage, scenario-had-ACT and right-censored. The at-risk file retains the same coordinates plus stage and event. Explicit linkage allows each summary or model row to be traced to the administered prompt and source response.

Stage-wise first-event analyses use the at-risk file rather than the complete master file. Because Group A has five stages whereas Groups B and C have ten, raw stage number is not interpreted as one common cross-group severity scale. Censored runs remain in the risk set through their final observed stage, and raw mean first-ACT stages are not compared mechanically across instruments with different observation windows. Figure 3 illustrates this separation with a worked example in which a post-ACT stage remains in the master record but is excluded from the first-event risk set.

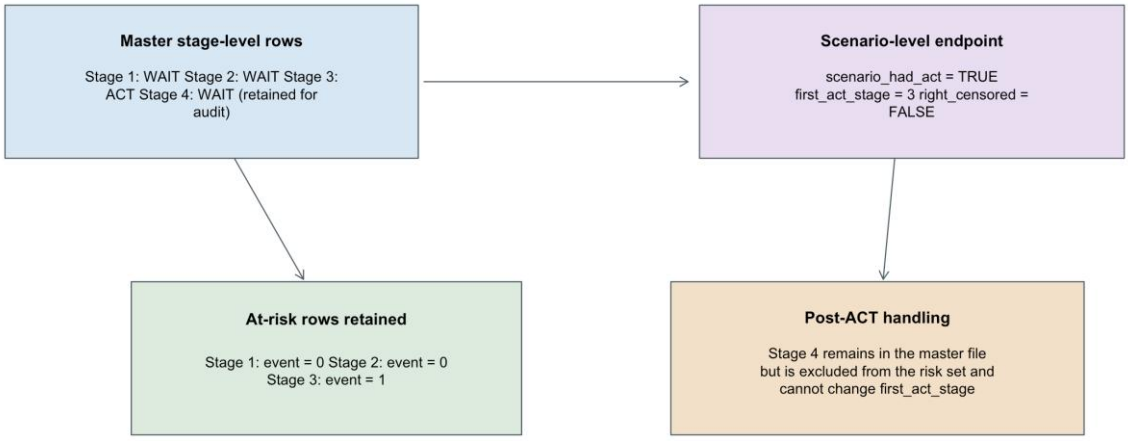

Figure 3: Worked endpoint example. A post-ACT stage remains in the master file for audit but is excluded from the first-event risk set.

2.9. Statistical analysis and analytical reproducibility

The revised primary inferential model is a binomial-logit discrete-time event-history model fitted to the 10,088 at-risk rows: event ~ stage * scenario_group + trajectory + prompt_role + response_mode + analysis_model_label. stage is numeric within scenario group; the stage-by-group interaction yields separate within-group linear trend summaries without treating raw stage as a common severity unit. The model includes 3,120 scenario-run clusters and uses HC1 cluster-robust covariance by scenario_run_id with 95% Wald intervals. This group-specific specification was added during revision to address the non-equivalent stage windows and was not pre-specified. The original pooled linear-stage model is retained in the public package as superseded analysis rather than described as pre-specified evidence for the revised specification.

A supplementary nonlinear sensitivity fitted event ~ (stage + I(stage^2)) * scenario_group + trajectory + prompt_role + response_mode + analysis_model_label with the same cluster-robust covariance. A saturated categorical stage-by-group model was not used because Groups A and B have zero Stage-1 events, creating complete separation for isolated Stage-1 indicators; the quadratic model tests curvature without adding a separate penalised estimation framework. The three group-specific quadratic components were tested jointly. Covariance sensitivities retained the fitted linear coefficients and recalculated uncertainty with broader clustering by scenario ID (24 clusters, t23 intervals) and

by analysis model (13 clusters, t12 intervals); the latter is treated only as a sensitivity because of the small number of model clusters.

All 10,088 at-risk rows are strict-compliant. The eight decision-valid non-strict responses and the single invalid response occur after an earlier ACT in their respective runs, so a strict-only first-event sensitivity is identical to the primary risk set. Group-specific odds ratios are interpreted only as compact within-group linear trend summaries; neither their magnitudes nor raw stage numbers are treated as directly comparable effect sizes across scenario groups.

The public release implements post-collection analytical reproducibility. Rscript 00_run_all.R reconstructs the first-ACT and at-risk datasets, validates prompt coordinates and release invariants, refits the linear and nonlinear models, regenerates manuscript tables and Figures 1–4, and checks numerical outputs against saved references. Newly authored scenarios reduce dependence on well-known benchmarks but do not prove absence of training exposure (Magar and Schwartz, 2022; Sainz et al., 2023); instrument dates and later wording/domain changes are therefore retained as versioned provenance rather than treated as a binary contamination label.

3 Results

3.1 Scale, parser performance and analytical reproducibility

The architecture was applied to thirteen canonical completed runs. Every model contributed the expected 2,000 stage-level records, producing 26,000 master rows and 3,120 scenario-runs. The at-risk file contains 10,088 rows: 3,097 first-ACT events and 6,991 WAIT observations; 23 runs are right-censored. The parser recovered a valid WAIT/ACT decision in 25,999 rows, of which 25,991 were strict-compliant. Eight decision-valid non-strict rows were recovered through the documented extracted-JSON pathway; all eight were Anthropic responses (seven Claude Sonnet 4.6 and one Claude Opus 4.8) and all occurred after an earlier ACT. The sole decision-invalid row was GLM-5.2, scenario C4B Stage 10, after first ACT at Stage 6, so none of the nine exceptions changes first-event inference.

Both derived datasets were regenerated exactly from the master record: the recreated scenario-level file matched all 3,120 deposited rows and the recreated at-risk file matched all 10,088 rows and event fields. The frozen instrument contained 200 ordered scenario-stage states crossed with five roles and two response modes, giving exactly 2,000 prompt coordinates. All 2,000 prompts were reconstructed exactly from deposited templates; no tested hidden construction-field name leaked into model-facing text; and all 20,000 non-missing master prompt_text values matched the frozen coordinate manifest. These are release and reconstruction checks, not evidence that the behavioural response pattern is normatively valid.

The complete one-command workflow for release v1.0.14 passed from a fresh extraction in the author's documented environment, regenerating every manuscript-facing table and Figures 1–4 and validating saved analytical outputs. Table 5 summarises the principal release invariants and parser categories. Thirty GLM-5.2 rows record more than one execution attempt, but attempts_used does not distinguish transport from semantic retries; these are therefore execution-attempt metadata rather than independent stochastic replicates.

Table 5: Release invariants and observed analytical-reproduction results

Validation domain / category	Required invariant	Observed demonstration result
Factorial completeness	Expected row count for every canonical model and stage coordinate.	13 models x 2,000 rows = 26,000; no missing canonical coordinates.
Decision-valid and strict-compliant	Exact WAIT/ACT decision; entire response matches the requested schema.	25,991 records (99.9654%); includes all 10,088 at-risk rows.
Decision-valid but non-strict	Exact WAIT/ACT decision recovered despite failure to meet the full schema.	8 records (0.0308%); all after first ACT, so no effect on first-event analysis.
Decision-invalid	No exact WAIT/ACT decision recovered under the fixed parser.	1 record (0.0038%); after first ACT, so no effect on first-event analysis.
Run reconstruction	One scenario-level row per model x scenario x role x mode.	3,120 unique scenario-runs.
Event uniqueness	Exactly one event per non-censored run and none for censored runs.	3,097 events; 23 right-censored runs.
Risk-set integrity	No at-risk row after an earlier first ACT.	10,088 correctly truncated at-risk rows.
Prompt reconstruction	One exact prompt per unique coordinate; templates exclude tested hidden fields.	2,000/2,000 exact prompt reconstructions; 20,000 retained master prompts matched; zero tested hidden-field leaks.
Cross-stage audit	Post-first-ACT rows retained for audit but excluded from first-event risk set.	652/3,097 ACT-containing runs had a later WAIT; 1,007/15,911 post-first-ACT rows were WAIT.

Note. Parser-category percentages use all 26,000 master records.

3.2. First-ACT distribution and provider provenance

Table 6 reports the complete first-ACT distribution by group and overall: 3,097 of 3,120 scenario-runs had an observed first ACT and 23 (0.74%) were right-censored. All censored runs were in Group A, whose observation window ends after five stages; Groups B and C have ten-stage instruments. Censoring rates are therefore not interpreted as substantively comparable across groups.

Table 6: First-ACT stage distribution by scenario group.

First-ACT stage / outcome	Group A, n (%)	Group B, n (%)	Group C, n (%)	Overall, n (%)
1	0 (0.00)	0 (0.00)	17 (1.63)	17 (0.54)
2	648 (62.31)	15 (1.44)	178 (17.12)	841 (26.96)
3	333 (32.02)	349 (33.56)	568 (54.62)	1,250 (40.06)
4	33 (3.17)	362 (34.81)	102 (9.81)	497 (15.93)
5	3 (0.29)	302 (29.04)	121 (11.63)	426 (13.65)
6	N/A	11 (1.06)	50 (4.81)	61 (1.96)
7	N/A	1 (0.10)	1 (0.10)	2 (0.06)
8	N/A	0 (0.00)	2 (0.19)	2 (0.06)
9	N/A	0 (0.00)	0 (0.00)	0 (0.00)
10	N/A	0 (0.00)	1 (0.10)	1 (0.03)
Right-censored	23 (2.21)	0 (0.00)	0 (0.00)	23 (0.74)

Note. Cells report n (%), with denominators of 1,040 scenario-runs per group and 3,120 overall, including censored runs. N/A denotes stages outside Group A's observation window. Percentages are rounded to two decimal places.

Requested identifiers, provider routes and returned identities were compared where serving metadata exposed an identity field. The three OpenAI canonical runs initially had blank returned-model fields in the merged file; identities were recovered from author-retained raw response bodies, written into the canonical analytical records and verified for all 6,000 OpenAI rows. The derived recovery table is public, but the underlying raw bodies are not generally deposited. Other returned-model identifiers are visible in the released master record, while exact row-level timing is public for OpenAI only. These route-specific provenance limits are documented rather than treated as random missingness.

3.3. Stage-wise instrument response

Figure 4 retains the same selected-stage display used in the original submission as a compact descriptive illustration of instrument responsiveness to escalating evidence, while Table 7 reports the complete conditional first-ACT profile for every observable stage, including numbers at risk and 95% Wilson intervals. The displayed stages were not re-selected during revision and are not the basis of the inferential analysis. The complete profiles are stage-dependent but non-monotone, and later estimates become increasingly uncertain as the risk sets shrink.

Group A rose sharply from near-zero early hazard to 84.9% at Stage 3 before declining to 11.5% by Stage 5. Group B increased from near zero to 96.2% at Stage 5, although only 12 runs remained at risk at Stage 6 and one at Stage 7. Group C showed the clearest non-monotonicity, with repeated increases and decreases across stages and only four, three, one and one runs remaining from Stages 7-10. The late-stage extremes therefore have wide uncertainty and should not be interpreted as stable response probabilities.

In the revised analysis, the group-by-stage model replaced the original common linear-stage coefficient. Adjusted odds ratios per additional stage within group were 12.591 (95% CI 8.481–18.694) in Group A, 10.346 (9.292–11.518) in Group B and 4.377 (3.815–5.022) in Group C. The stage-by-group interaction was jointly non-zero (robust Wald chi-square(2) = 116.946, p = 4.03e-26), confirming that one pooled stage coefficient is inappropriate. A quadratic stage-by-group sensitivity also showed strong nonlinearity (robust Wald chi-square(3) = 305.587, p = 6.14e-66). These estimates are compact within-group trend and shape summaries; the odds-ratio magnitudes are not interpreted as substantively comparable across the three non-equivalent instruments. Covariance sensitivities clustered by scenario ID or analysis model retained positive within-group trends, while the model-cluster result is treated only as a sensitivity because there are thirteen model clusters.

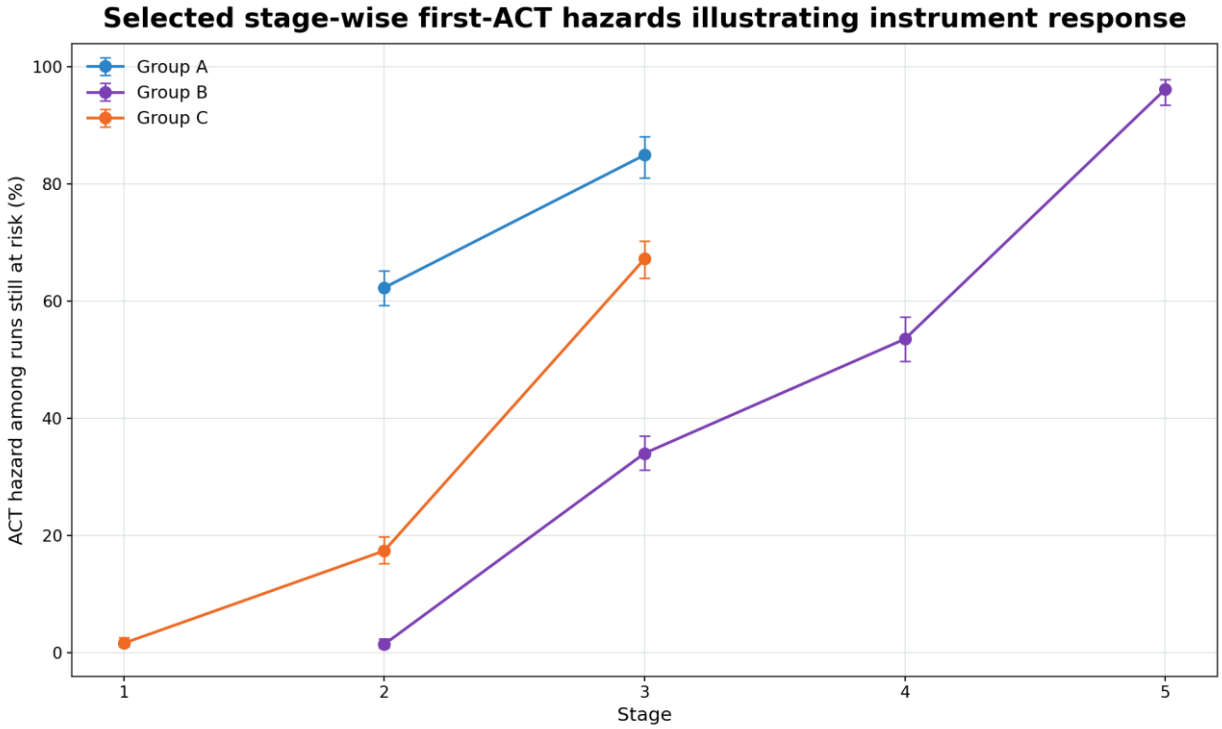

Figure 4: Selected stage-wise first-ACT hazards illustrating instrument response. The displayed stages are unchanged from the original submission (Group A: Stages 2-3; Group B: Stages 2-5; Group C: Stages 1-3), where they were used solely as a compact descriptive illustration of instrument responsiveness to escalating evidence. Points show conditional hazards among runs still at risk; vertical bars show 95% Wilson intervals. The *displayed subset is not used to estimate or test stage effects; complete stage-wise hazards, numbers at risk and 95% confidence intervals are reported in Table 7 and the reproducibility materials, and the inferential analyses use the full 10,088-row at-risk dataset.*

Table 7: Complete conditional first-ACT hazards by scenario group.

Group	Stage	n at risk	n ACT	Conditional ACT hazard (%)	95% Wilson CI
A	1	1,040	0	0.0	0.0–0.4
A	2	1,040	648	62.3	59.3–65.2
A	3	392	333	84.9	81.1–88.1
A	4	59	33	55.9	43.3–67.8
A	5	26	3	11.5	4.0–29.0
B	1	1,040	0	0.0	0.0–0.4
B	2	1,040	15	1.4	0.9–2.4
B	3	1,025	349	34.0	31.2–37.0
B	4	676	362	53.6	49.8–57.3
B	5	314	302	96.2	93.4–97.8
B	6	12	11	91.7	64.6–98.5
B	7	1	1	100.0	20.7–100.0
C	1	1,040	17	1.6	1.0–2.6
C	2	1,023	178	17.4	15.2–19.8
C	3	845	568	67.2	64.0–70.3
C	4	277	102	36.8	31.4–42.6
C	5	175	121	69.1	61.9–75.5
C	6	54	50	92.6	82.4–97.1
C	7	4	1	25.0	4.6–69.9
C	8	3	2	66.7	20.8–93.9
C	9	1	0	0.0	0.0–79.3
C	10	1	1	100.0	20.7–100.0

Note. Hazard is n ACT / n at risk at each observable stage. Intervals are 95% Wilson intervals. No Group B run remained at risk after Stage 7. Figure 4 retains the original-submission selected-stage descriptive display; the complete profile reported here is the basis for descriptive interpretation.

3.4. Post-First-ACT consistency audit

Because the full manifest continued after first ACT, later independent calls provide an audit unavailable under physical stopping. Of 3,097 ACT-containing runs, 3,093 had at least one later decision-valid stage and 652 (21.05%) later produced at least one WAIT. Across 15,911 decision-valid post-first-ACT rows, 1,007 (6.33%) were WAIT; 277 of 3,093 eligible runs (8.96%) returned WAIT at the immediately following stage. Among affected runs, the mean number of later WAIT rows was 1.54, the median was 1 and the maximum was 4; the median distance from first ACT to the first later WAIT was two stages (maximum eight). These reversals do not establish that first-ACT is a uniquely correct behavioural endpoint. They show that later independent judgements can differ after an observed ACT, making retained post-event outputs diagnostically useful. Their exclusion from the at-risk dataset follows separately from the definition of the first-event estimand. Possible explanations include stochastic response variation, decision-boundary sensitivity, compact-persistence interpretation and changing emphasis across later evidence; the present data do not identify which mechanism generated a given reversal.

4 Discussion

4.1. Integrated contribution

The framework does not claim novelty for sequential evaluation, stopping rules, discrete-time event-history analysis, or provenance recording individually. The wide-to-long first-action/censoring and at-risk transformation follows prior stopping-rule reconstruction work (Kolesnikov, 2026b). The contribution is its integration and extension for mutable served LLM systems: cumulative evidence is separated from conversational self-history by independent current-state calls; prompt/parser policy is frozen and reconstructable; provider route and requested/returned identity become part of the analytical provenance; and post-first-ACT model outputs are deliberately retained for audit while inherited first-event logic excludes them from the risk set.

The reconstruction checks establish analytical/post-collection reproducibility rather than exact response-level reproducibility. The first-ACT and at-risk files, prompt coordinates and manuscript outputs can be regenerated from deposited records without provider access, but historical provider responses cannot be recreated independently of mutable endpoints, wrappers and serving infrastructure. Likewise, the stage-wise hazard profiles are an instrument-response check rather than validation that a particular intervention boundary is correct. This narrower claim is important because each empirical first-ACT is one realised response path under its recorded serving conditions.

4.2. Comparison with alternative evaluation designs

Table 8 positions the architecture against representative static, multi-turn, agentic, stateful-tool and stopping-rule approaches across first-event/censoring, self-history exposure, post-event audit, reproducibility and served-system provenance.

Table 8: Comparison with representative sequential and interactive evaluation approaches.

Approach	Primary endpoint	First-event / censoring	Self-history exposure	Post-event audit	Reproducibility basis	Served-system provenance
HELM (Liang et al., 2023)	Broad multi-metric performance	No	Not focal	No first-event construct	Standardised scenarios/metrics	Not route-specific
MT-Eval (Kwan et al., 2024)	Multi-turn dialogue capability	No	Earlier turns retained	Not a first-event audit	Published tasks/procedure	Not central
MT-Bench-101 (Bai et al., 2024)	Fine-grained multi-turn dialogue	No	Earlier turns retained	Not a first-event audit	Tasks/taxonomy/scoring	Not central
AgentBench (Liu et al., 2024)	Interactive agent task performance	No	History integral	No first-event separation	Benchmark environments/tasks	Not central
tau-bench (Yao et al., 2025)	Final database state; pass^k reliability	No	Persistent user-agent-tool history	No first-event separation	Tools/policies/goals/repeated trials	Not central
ToolSandbox (Lu et al., 2025)	Stateful tool trajectories/milestones	No	Conversation/state retained	Trajectory audit, not first-event	Stateful sandbox/tasks/milestones	Not central
Human stopping-rule pipeline (Kolesnikov, 2026b)	First-action/censoring + at-risk reconstruction	Yes	No LLM self-history problem	No: survey stops after action	Configured reconstruction/tests	No provider/model route
Present architecture	One realised first-ACT/censoring path + audit	Yes	Prior model outputs excluded by construction	Yes: later model outputs retained in master	Frozen prompt/parser + deterministic reconstruction	Provider route, identity and collection window

The comparison is deliberately about evaluation targets rather than superiority. HELM emphasises breadth across standardised metrics; MT-Eval and MT-Bench-101 evaluate persistent multi-turn dialogue; AgentBench, tau-bench and ToolSandbox evaluate interactive or stateful task performance and reliability; and the prior human stopping-rule pipeline addresses reproducible first-action reconstruction outside LLMs. The present architecture targets the narrower problem of representing and reconstructing when a served model first recommends action under accumulating evidence while preserving route-level provenance and a post-event audit trail.

4.3. Reuse, interpretation and adaptation

The three primary files support different analytical tasks. 01_data/d2_master_stage_level_13_models.csv is the authoritative source for parser audits, provenance, post-ACT consistency checks and alternative endpoint definitions; 01_data/d2_first_act_scenario_level_13_models.csv supports one-row-per-run summaries and censoring analyses; and 01_data/d2_stage_level_at_risk_hazard_dataset.csv supports first-event modelling. The 2,000-coordinate prompt manifest in 02_instrument_and_prompts/ supports exact prompt reconstruction. Treating these files as interchangeable would undo the architecture's central distinction between source record and analytical risk set.

Interpretation should remain time-stamped and route-specific. A later endpoint carrying a similar family name is not automatically the same served system, and prompt-role results are effects of the deposited wordings rather than universal constructs such as safety or expertise. Group A has five stages whereas Groups B and C have ten; raw stage positions therefore are not common severity units, and the different group-specific odds-ratio magnitudes are not given a substantive between-group interpretation.

The demonstrated scope is text-only served LLMs, one binary WAIT/ACT event definition, English-language prompts, independent completed-response calls and provider-specific API routes. Reasoning models can use the same observable first-event logic without treating hidden reasoning traces as part of the endpoint. Multimodal systems would require versioned media inputs and documented preprocessing. Tool-using agents require a pre-defined qualifying action. The present framework can evaluate its first occurrence; modelling competing action types, repeated interventions or subsequent reversals would require an appropriate extension beyond the current binary first-event model. Streaming APIs require a pre-specified choice between completed-response and token/chunk-time events. Non-English prompts require a separately frozen language-specific instrument and equivalence checks rather than cosmetic translation.

Serving changes should define new evaluation versions: a changed endpoint, provider wrapper, returned model identity or route should create a new time-stamped cohort rather than be silently pooled with the historical run. A concrete non-LLM application already exists in human sequential-decision data: Kolesnikov (2026b) used the same broad first-action/censoring and at-risk reconstruction logic for stopping-rule survey exports. Human scenarios physically stop after action, whereas the LLM design can deliberately retain later independent outputs and must additionally preserve mutable model/route provenance.

Deployment-oriented use requires the ACT definition to match the action available in that system. Seeking advice, locking an account, clinical escalation and issuing a public warning have different costs and should not be pooled under one unexamined label. Claims about an optimal threshold require external expert judgement, outcome data or an explicit cost function. The present framework supplies an auditable measurement of realised behavioural timing, not the normative standard against which that timing should be judged.

4.4. Limitations

- The demonstration uses simulated written scenarios. It does not validate clinical correctness, prevent financial loss, detect live cybersecurity incidents or predict user response. The study provides no independent validation of the stage-severity ordering or evidence escalation. Hidden construction weights are not operational severity scales, and an observed first-ACT stage is not automatically the normatively correct intervention point.
- Cumulative persistence is sometimes expressed compactly rather than by repeating every earlier sentence verbatim. Eight decision-valid non-strict Anthropic responses explicitly commented on omitted earlier detail, although all occurred after first ACT. This does not alter endpoint reconstruction but shows that compact persistence was not interpreted identically in every response; a fully enumerated cumulative instrument is a separate robustness condition.
- Independent current-state calls prevent earlier model outputs from entering later inputs by construction, but the original cohort did not include a persistent-conversation comparison. A stratified comparison was specified during revision but was not executed because the original endpoints were no longer uniformly available to the author under the original routes; substituting current models would create a different cohort rather than reproduce the historical served systems.
- The five prompt roles use one wording each and the empirical instrument is English-language. The results therefore support those frozen wordings, not universal policy constructs or cross-language equivalence. Multimodal, tool-using, streaming and non-English adaptations are proposed extensions rather than demonstrated capabilities.
- Each reported first-ACT stage is one realised response path under a recorded route, time window, prompt condition and provider-standard sampling condition. The study uses one primary response per frozen coordinate and therefore does not estimate within-condition stochastic stability or a latent invariant intervention threshold.
- The cohort is time-stamped and route-dependent. Provider wrappers, safety layers, endpoint aliases and serving configurations may change. The deposited workflow is analytically reproducible after collection, but exact historical provider-response reproduction is not guaranteed; later collections should be treated as new versioned cohorts.
- Group A has five stages whereas Groups B and C have ten. Raw or mechanically normalised stage positions are not common severity units. The revised group-specific linear odds ratios are within-group trend summaries, and complete hazards plus the nonlinear sensitivity are required to interpret stage shape.
- Because 3,097 of 3,120 runs eventually produced ACT, the demonstration primarily discriminates when models act rather than whether they ever act. The present architecture is also a single first-event design. Later repetition or reversal does not invalidate analysis of the first qualifying action; modelling competing action types, repeated interventions or subsequent reversals requires an appropriate extension and observation window.

5 Conclusion

The architecture makes first-intervention timing analytically reconstructable from a provenance-preserving source record. Across the 26,000-record cohort, exact regeneration of endpoint files and prompt coordinates demonstrates post-collection analytical reproducibility, while the complete stage-wise profiles provide a descriptive check that the frozen instrument elicits stage-dependent responses. The framework does not identify a normatively correct intervention stage or a stable latent threshold of a served model, and it cannot recreate historical provider responses independently of mutable serving infrastructure.

Its contribution is narrower: one realised first-action path can be audited as a first-event process while post-event model outputs remain available for diagnosis, and the same source-record/first-event/at-risk separation can be reused for new, explicitly time-stamped cohorts. Broader deployment requires setting-specific event definitions, provenance fields and validation rather than assuming that the demonstrated text-only implementation transfers unchanged.

Data availability statement

The released data, frozen prompt materials, codebooks, audit records, reconstruction scripts and manuscript outputs are openly available in the Open Science Framework repository “A Reproducible, Provenance-Preserving Architecture for Sequential Intervention-Threshold Evaluation in Large Language Models: Reproducibility Materials” (Kolesnikov, 2026c), DOI: https://doi.org/10.17605/OSF.IO/T58R9. The manuscript corresponds to release v1.0.14 of the reproducibility package. The deposited materials support reconstruction of the derived datasets, reproduction of the analyses and regeneration of the reported tables and figures. Datasets, scenario texts, prompt templates, documentation, tables and figures are licensed under CC BY 4.0; software code is licensed under the MIT License. Some raw provider response files are retained locally and are not publicly deposited.

Ethics and consent

The LLM evaluation used hypothetical, author-created text scenarios. No new human participants, animal experiments or social-media data were involved. Group A reuses author-created text stimuli from the anonymous human sequential-decision instrument described in Kolesnikov (2026a); no human participant data are included in the released LLM datasets or analysed in this article. The released model-response data contain no personal identifiers, clinical records or user-level behavioural data.

Author contributions

Alex Kolesnikov: Conceptualisation; Methodology; Software; Validation; Formal analysis; Investigation; Resources; Data curation; Writing - original draft; Writing - review and editing; Visualisation; Project administration.

Acknowledgements

The author thanks Ellie Fassman for her guidance as a project mentor and for reviewing the results and overall structure of this paper.

Use of AI-assisted tools

During preparation of this manuscript, the author used ChatGPT to assist with language editing, manuscript organisation, and clarity. The tool did not generate the study data, determine the reported results, or make scientific conclusions. The author reviewed all AI-assisted suggestions and takes full responsibility for the originality, accuracy, and integrity of the manuscript.

Funding information

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

References

[1] Allison, P.D. (1982) ‘Discrete-time methods for the analysis of event histories’, Sociological Methodology, 13, pp. 61–98. Available at: https://doi.org/10.2307/270718.

[2] Aloqalaa, M., Soiland-Reyes, S. and Goble, C. (2026) ‘A structured examination of reproducibility: a case study for HTS using the PRIMAD model and BioCompute Object’, Data Science Journal, 25, article 24. Available at: https://doi.org/10.5334/dsj-2026-024.

[3] Bai, G., Liu, J., Bu, X. et al. (2024) ‘MT-Bench-101: a fine-grained benchmark for evaluating large language models in multi-turn dialogues’, Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 7421–7454. Available at: https://doi.org/10.18653/v1/2024.acl-long.401.

[4] Edmunds, S.C., Nogoy, N., Lan, Q. et al. (2026) ‘Integrating machine learning standards in disseminating machine learning research’, Data Science Journal, 25, article 1. Available at: https://doi.org/10.5334/dsj-2026-001.

[5] Gallifant, J., Afshar, M., Ameen, S. et al. (2025) ‘The TRIPOD-LLM reporting guideline for studies using large language models’, Nature Medicine, 31(1), pp. 60–69. Available at: https://doi.org/10.1038/s41591-024-03425-5.

[6] Green, D.M. and Swets, J.A. (1966) Signal detection theory and psychophysics. New York: Wiley.

[7] Kolesnikov, A. (2026a) ‘A color-coded urgency bar and intervention thresholds in sequential health decisions’, Journal of High School Science, 10(3), pp. 394–413. Available at: https://doi.org/10.64336/001c.166227.

[8] Kolesnikov, A. (2026b) ‘A reproducible computational pipeline for modelling sequential decision thresholds from stopping-rule data’, International Journal of Secondary Computing and Applications Research, 3(3), pp. 14–20. Available at: https://doi.org/10.67149/yhjs2024.5/3s7gdh1q.

[9] Kolesnikov, A. (2026c) A Reproducible, Provenance-Preserving Architecture for Sequential Intervention-Threshold Evaluation in Large Language Models: Reproducibility Materials. Open Science Framework. Available at: https://doi.org/10.17605/OSF.IO/T58R9.

[10] Kwan, W.-C., Zeng, X., Jiang, Y. et al. (2024) ‘MT-Eval: a multi-turn capabilities evaluation benchmark for large language models’, Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 20153–20177. Available at: https://doi.org/10.18653/v1/2024.emnlp-main.1124.

[11] Liang, P., Bommasani, R., Lee, T. et al. (2023) ‘Holistic evaluation of language models’, Transactions on Machine Learning Research. Available at: https://openreview.net/forum?id=iO4LZibEqW (Accessed: 16 September 2026).

[12] Liu, X., Yu, H., Zhang, H. et al. (2024) ‘AgentBench: evaluating LLMs as agents’, The Twelfth International Conference on Learning Representations (ICLR 2024). Available at: https://openreview.net/forum?id=zAdUB0aCTQ (Accessed: 16 September 2026).

[13] Lu, J., Holleis, T., Zhang, Y. et al. (2025) ‘ToolSandbox: a stateful, conversational, interactive evaluation benchmark for LLM tool use capabilities’, Findings of the Association for Computational Linguistics: NAACL 2025, pp. 1160–1183. Available at: https://doi.org/10.18653/v1/2025.findings-naacl.65.

[14] Magar, I. and Schwartz, R. (2022) ‘Data contamination: from memorization to exploitation’, Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 157–165. Available at: https://doi.org/10.18653/v1/2022.acl-short.18.

[15] Royal College of Physicians (2017) National Early Warning Score (NEWS) 2: standardising the assessment of acute-illness severity in the NHS. Updated report of a working party. London: Royal College of Physicians. Available at: https://www.rcp.ac.uk/resources/national-early-warning-score-news-2/ (Accessed: 16 September 2026).

[16] Sainz, O., Campos, J.A., García-Ferrero, I. et al. (2023) ‘NLP evaluation in trouble: on the need to

measure LLM data contamination for each benchmark', Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 10776–10787. Available at: https://doi.org/10.18653/v1/2023.findings-emnlp.722.

[17] Sclar, M., Choi, Y., Tsvetkov, Y. et al. (2024) 'Quantifying language models' sensitivity to spurious features in prompt design or: how I learned to start worrying about prompt formatting', The Twelfth International Conference on Learning Representations (ICLR 2024). Available at: https://openreview.net/forum?id=RIu5lyNXjT (Accessed: 16 September 2026).

[18] Singer, J.D. and Willett, J.B. (2003) Applied longitudinal data analysis: modeling change and event occurrence. New York: Oxford University Press. Available at: https://doi.org/10.1093/acprof:oso/9780195152968.001.0001.

[19] Srivastava, A., Rastogi, A., Rao, A. et al. (2023) 'Beyond the imitation game: quantifying and extrapolating the capabilities of language models', Transactions on Machine Learning Research. Available at: https://openreview.net/forum?id=uyTL5Bvosj (Accessed: 16 September 2026).

[20] Wald, A. (1945) 'Sequential tests of statistical hypotheses', Annals of Mathematical Statistics, 16, pp. 117–186. Available at: https://doi.org/10.1214/aoms/1177731118.

[21] Yao, S., Shinn, N., Razavi, P. et al. (2025) 'tau-bench: a benchmark for tool-agent-user interaction in real-world domains', The Thirteenth International Conference on Learning Representations (ICLR 2025). Available at: https://openreview.net/forum?id=roNSXZpUDN (Accessed: 16 September 2026).