Building trustworthy enterprise benchmarks and synthetic training data

Boundary design, recipe-driven generation and frozen training packs

Mark J Sykes, AI One Labs | 6 November 2026

Executive summary

A useful enterprise dataset teaches the distinction that changes the next action. Reviewing a code change and implementing it may involve the same repository, people and vocabulary. The data must make the requested operation clear, while testing whether the model can find it in realistic context.

HISTORICAL BENCHMARK	NEW FROZEN CORPUS	PRODUCTION SPEND
1,974	6,497	$156.93
50 workflow categories	Train, validation and test	Provider-reconciled usage

This paper documents two complementary assets: a difficult, reusable benchmark and a separate corpus for specialising Clef. The benchmark preserves related scenario families and explicit workflow boundaries. The training programme uses declared coverage, eligible model routes, blind review, family exclusions and a reproducible export.

Synthetic Data Studio helped make recipes, generation progress, public inputs and private audit records inspectable. Frozen packs carried the task and its evidence to an external trainer. The original benchmark predates Studio; its lineage and evaluation-only status were preserved.

The approach produced the data behind Paper 2's measured reduction in historical Clef errors. Its strongest contribution is a repeatable method for constructing and auditing task-specific evidence. Reference labels are model-authored and reviewed automatically, not individually adjudicated human ground truth.

Benchmark design: pp. 3–9. Training recipes, routes and review: pp. 10–15. Packs and encoding: pp. 16–17. Economics: p. 18. Taxonomy and reproducibility: pp. 19–20. Screenshots show retained application records inspected on 6 November 2026.

Making the research process inspectable

The practical problem was to turn a fixed enterprise decision task into evidence that could survive interrupted generation, review disagreements and transfer to a separate training machine. Studio provided a useful workbench for inspecting the recipe, coverage, examples and accounting; campaign tooling handled the study-specific exclusions and final pack assembly.

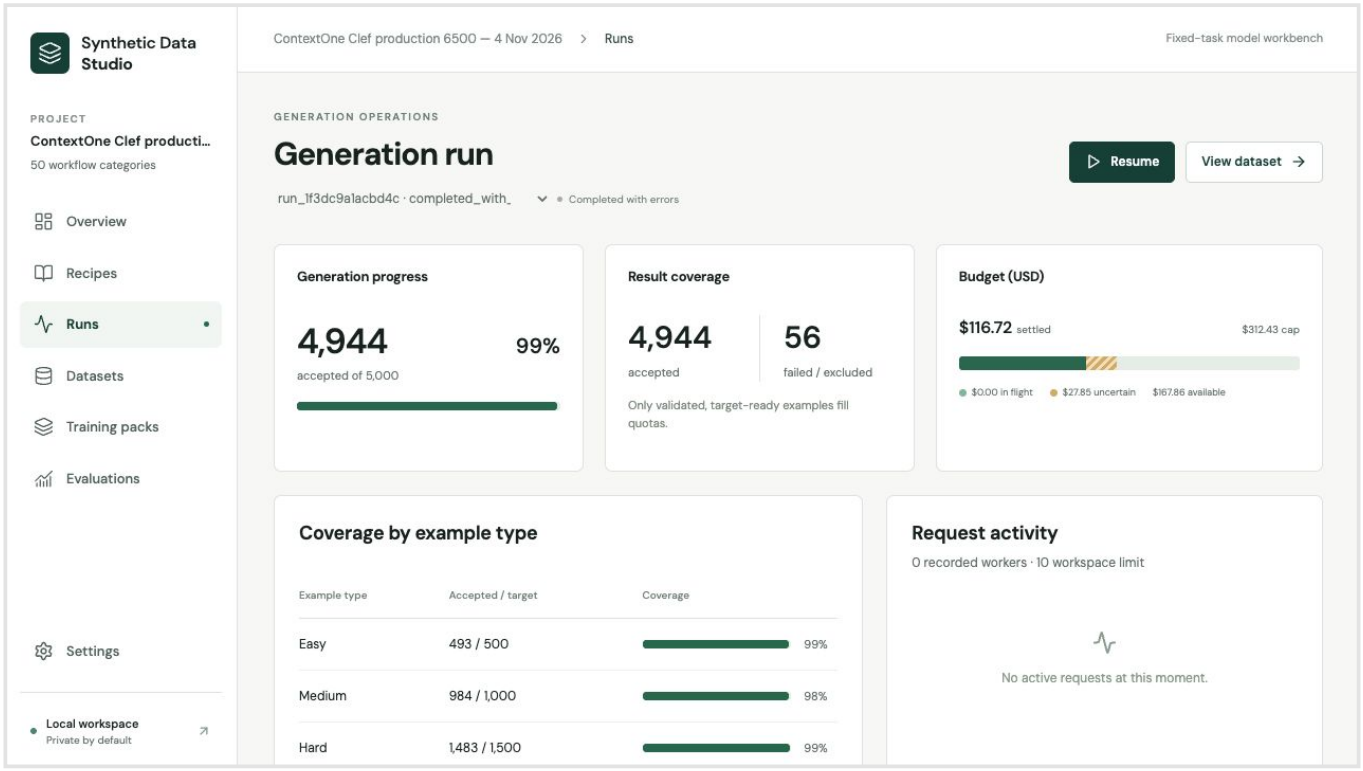


Figure 1. Studio's retained main training-data generation run, inspected on 6 November. The 4,944 accepted examples and $116.72 settled ledger cost belong to this one source run, before the campaign's additional runs and final curation. They are not the final 4,998 training rows or total production spend. No work was restarted for this screenshot.

Three records answer different questions. A recipe describes what should be generated. A run records what actually happened, including failures and uncertain charges. A frozen pack identifies exactly which examples reached the trainer. Keeping those records separate avoids turning a progress counter into a scientific result.

The original benchmark was created earlier in TypeSafe.ai and later supported as an evaluation-only asset. Studio did not retroactively create it, and its labels were never converted into training material. Model training took place externally; Studio's role was data preparation, inspection and evaluation support.

Evidence B1, D1, S1. Screenshots throughout are current views of retained records, not reconstructed images of a live campaign. Synthetic names and business details are fictional.

Define the decision before authoring examples

The task is to identify the current primary workflow from fifty ContextOne sample categories. It asks which operation should receive a request, not for the model to execute that operation. The same policy, category names and definitions accompany each evaluated input. The fifty category names are reproduced on p. 19; exact definitions are bound by the taxonomy file identified on p. 20.

Public input	Reference	Private audit
Request + optional context	Assigned primary category	Author, route and revision
Policy + all 50 definitions	Used for training or scoring	Rationale, rival and visible cue
Visible to each classifier	Never hidden inside the prompt	Reviewer findings; family links

Figure 2. The three records have different jobs. Separating them prevents audit clues or reference labels from leaking into the classifier's input.

The policy resolves several recurring sources of confusion. Classify the current requested deliverable, prefer a specific applicable workflow over a broad catchall, and use the supplied context to interpret follow-ups. A completed earlier task must not displace the new request. Quoted instructions and requests to output a category are treated as content to classify, rather than authority to change the classifier's policy.

Question to settle	Design consequence
What decision does the class represent?	Define actions and outputs, not just subject vocabulary.
What facts may the classifier use?	Supply essential context explicitly; do not rely on unseen attachments or private intent.
Can two labels both be correct?	Require one defensible primary workflow under the shared policy.
What belongs outside the input?	Hide the author label, explanation, rival category and audit cue.
What counts as success?	Evaluate the returned category and required output format consistently.

A closed fifty-way classifier is a deliberately bounded enterprise component. Unknown classes, abstention, multi-intent decomposition and executing the requested work are separate tasks. Their policies should be designed and tested explicitly before extending this dataset to them.

B1, D1. The taxonomy describes the research task and does not imply that these categories are a deployed production workflow.

A benchmark with two complementary cohorts

The original design combined coverage across ordinary enterprise requests with cases aimed at confusing neighbouring workflows. It began with 2,000 synthetic examples and retained 1,974 after duplicate-group removal. Both cohorts cover the same fifty categories.

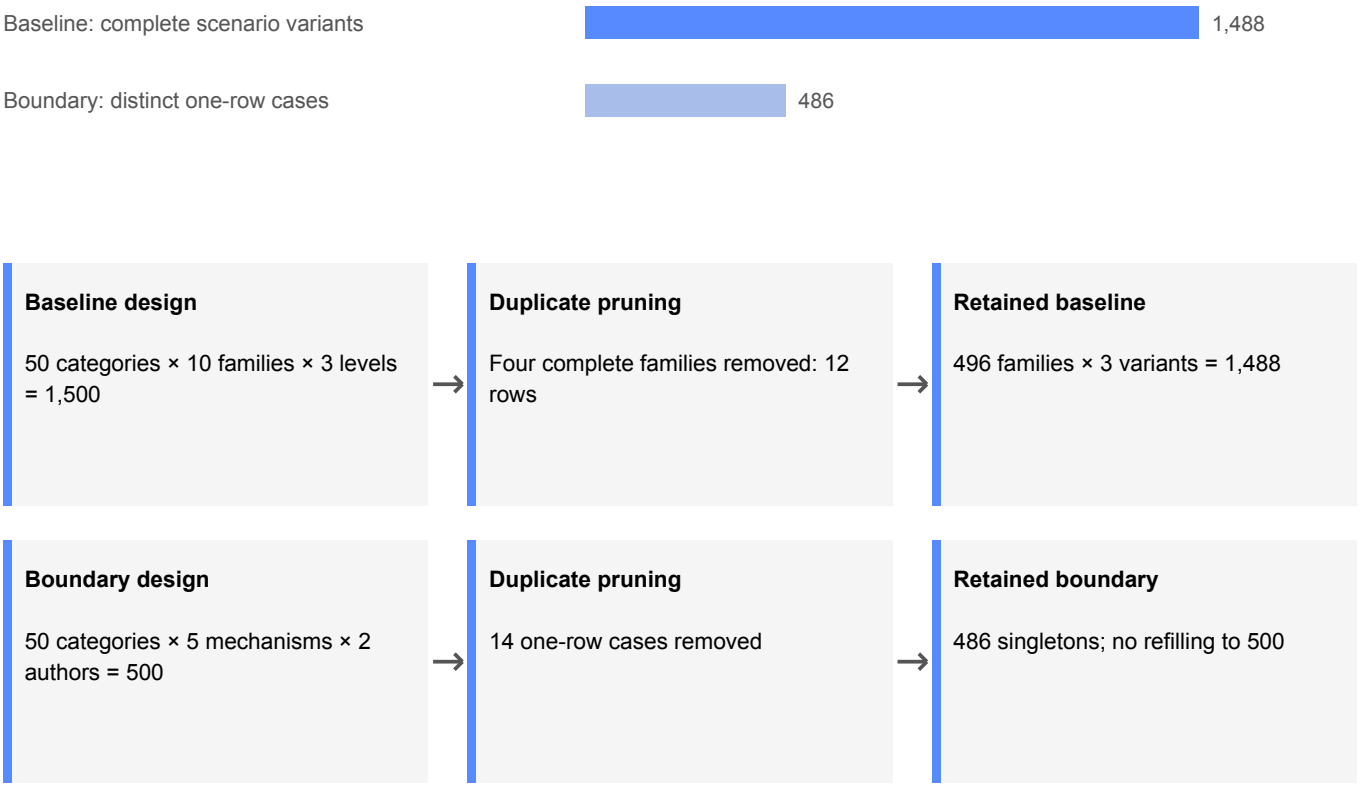


Figure 3. The final cohort is 1,974 rows across 982 explicit scenario components. The two design strata should remain identifiable in analysis.

The baseline contributes 75.38% of scored rows and the boundary cohort 24.62%. Forty-five categories retain forty examples each; categories 12 and 26 have 34, categories 8 and 35 have 35, and category 50 has 36. Reporting macro F1 alongside accuracy reduces the influence of these small class-count differences.

The benchmark is compact enough to run repeatedly but more structured than a random list of prompts. It preserves author, family, level and mechanism identities so the same total score can be examined by scenario type. Its public requests and context have a median combined length of 75 words, a 95th percentile of 138 and a maximum of 204; 1,895 rows contain context and 79 do not.

B1. Rows are not independent scenarios: the 496 three-variant families and 486 singletons form 982 components. Resample complete components when estimating historical comparison uncertainty.

Scenario families make coverage deliberate

For each category, the initial plan assigned ten scenario families with three intended levels per family. Five families went to DeepSeek V4 Pro 0813 and five to Kimi K3. The author received the full taxonomy, task policy, assigned category, exact slots, style suggestions and a variation seed. It invented fictional scenario content within that structure.

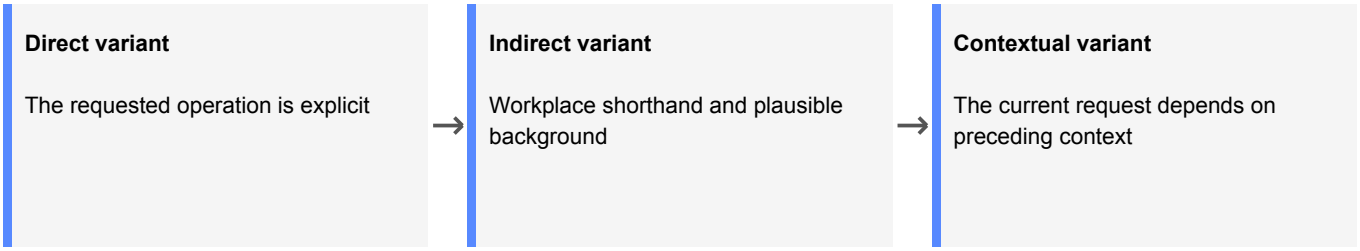


Figure 4. A baseline family contains three related variants. This is a schematic of the authored levels, not a verbatim family from the benchmark.

Style suggestions included informal messages, formal requests, contextual follow-ups, competing background goals and business shorthand. They were prompts for diversity, not hand-written scenario briefs. Fixed seeds controlled planning, review grouping and ordering; preserved outputs, rather than a promise of deterministic remote generation, make the experiment reproducible.

Property	Recorded benchmark method
Author routes	DeepSeek V4 Pro 0813 on DeepInfra FP8; Kimi K3 on DeepInfra BF16, through OpenRouter
Reasoning / output allowance	Max reasoning; 16,384 output tokens for both authors
Retained author totals	DeepSeek 974; Kimi 1,000
Baseline family structure	496 complete families; one direct, indirect and contextual variant each
Public input	Request and optional context; no explicit reference IDs or category-answer hints

An authored “hard” example is not necessarily hard for a model. The initial 1,500 baseline cross-reviews produced only 34 category disagreements, and the nominally hardest level produced fewer than the middle level. Some prompts made the answer obvious by ruling out alternatives. That evidence motivated a separate boundary cohort before scored competitor testing; it was not constructed from Clef’s later errors.

B1. These benchmark routes differ from the later training-data routes on p. 11. The benchmark was generated through the earlier TypeSafe.ai pipeline, not the later Studio campaign.

Hard boundaries need visible evidence

A difficult routing example places plausible alternatives close together while leaving a decisive cue in the public input. Removing that cue would create an underspecified question rather than a stronger test. The boundary cohort uses five mechanisms, each planned across all fifty categories and both authors.

Mechanism	Retained	What must be resolved
Contextual follow-up	98	The short request acquires its referent and operation from context.
Lexical decoy	97	Background vocabulary suggests a different workflow from the requested action.
Deliverable versus topic	97	The source material and requested artefact belong to different workflows.
Workflow stage	97	Deciding, implementing, recording and reviewing must be distinguished.
Scope change	97	The latest request supersedes the operation implied by earlier context.

Two requests about code, two different actions

Contextual follow-up. “Can you give it another look before standup at 10?” The supplied conversation says that retry-logic changes for a webhook timeout fix were pushed as PR #412. The reference is **Code review workflow (46)**. The requested review concerns a change that already exists; the short request alone would be insufficient.

Lexical decoy. “The reviewer flagged three issues in the comment thread. Apply the changes to the feature branch, update the affected tests, and push.” The reference is **Code change implementation (45)**. The word “reviewer” identifies the source of the instructions, while the requested actions require implementation.

The boundary author privately recorded a nearest alternative, the decisive visible cue and why the alternative loses. Those fields supported inspection but were hidden from blind reviewers and evaluated classifiers. Public text was prohibited from giving category names, elimination lists or a final sentence that simply restated the taxonomy definition.

B1. Actual fictional examples b-c46-s02 and b-c45-s03. Context in the first case is abridged; request text is quoted verbatim. Each example’s blind reviewer agreed with its author-assigned reference.

Topic, stage and scope change the destination

A meeting can produce an audit record

“From the CAB meeting and RFC-441, produce the formal decision record: timeline, approvers, conditions, and evidence IDs.” The reference is **Audit trail generation (19)**, rather than Meeting-to-actions workflow. The source is a meeting, but the requested output is a formal evidence record. A model that routes on the word “meeting” misses the deliverable.

An approved credit can become a record update

“Finance signed off. Post the £312.50 credit memo against INV-8892 in the billing system and link it to the customer's account.” The context records that the service credit was judged warranted, the amount calculated and finance approval received. The reference is **System-of-record update workflow (13)**. The current operation is posting an authorised record, rather than deciding or processing refund eligibility.

A CRM list can become an enrichment task

“Pull employee count and current tooling for each, assign a lead grade, and reorder the list for follow-up.” The context says that 55 event leads are already in Salesforce with source and campaign fields set. The reference is **Lead enrichment and scoring (31)**. The new request adds research and grading; the fact that the records live in a CRM does not make the task routine record maintenance.

Design rule	Failure it is intended to expose
Separate source from deliverable	Routing a meeting-derived audit record as meeting actions
Make chronology visible	Treating a completed eligibility decision as the current task
Preserve the latest request	Continuing CRM-entry work after the user asks for enrichment
Withhold private author cues	Scoring an explanation-assisted task that users never actually submit

These are checks of the shared classification policy, not legal, financial or software-execution advice. The requested work is never performed during the benchmark. The five illustrations on these two pages were chosen to explain the mechanisms and do not constitute a representative error sample.

B1. Actual fictional examples b-c19-s05, b-c13-s08 and b-c31-s09. Requests quoted verbatim; contexts abridged. Each example’s blind reviewer agreed with its assigned reference and found the context sufficient.

Blind review supplies a second interpretation

Each benchmark author's examples went to the other model in a fresh request. The reviewer saw the policy, taxonomy and public input under opaque identifiers. It did not see the assigned category, intended level, author explanation, nominated alternative, mechanism or decisive cue. Batches contained at most ten rows and never two variants from the same baseline family.

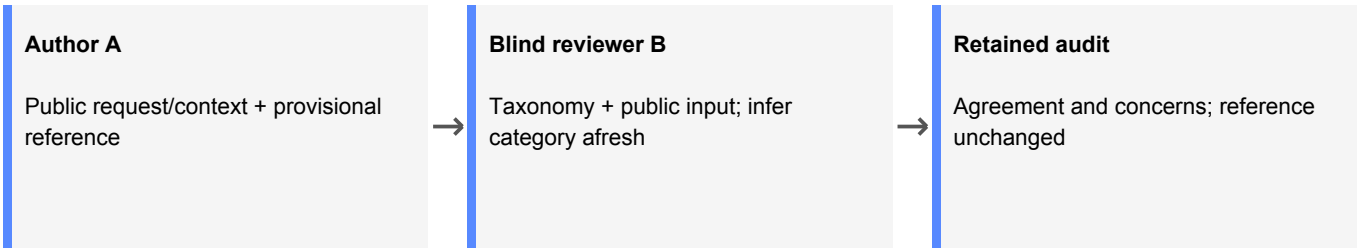


Figure 5. Original benchmark cross-review. DeepSeek reviewed Kimi and Kimi reviewed DeepSeek; fresh requests hid author metadata and reference labels.

Advisory finding	Baseline, 1,488	Boundary, 486	Total
Category disagreement	34	33	67
Difficulty disagreement	222	404	626
Possible ambiguity	4	21	25
Possibly insufficient context	0	1	1

Findings overlap: 676 rows have at least one advisory flag. Category disagreement was 2.28% in the baseline and 6.79% in the boundary cohort. This shows that the second author more often found a competing interpretation in boundary cases; it does not establish that the reference was wrong or assign a calibrated difficulty score.

The reference remained the author-assigned provisional label. Reviewer disagreement was advisory, avoiding an automatic filter that would retain only consensus-easy cases. Blocking structural checks and bulk approval completed the freeze; the benchmark did not receive individual human adjudication. Accuracy is therefore agreement with a fixed, auditable reference set, not an assertion of error-free human ground truth.

For an enterprise using this method, disagreement should remain visible. If the business cannot resolve a boundary from its own policy, adding more examples will not settle which action it intended. Clarify the policy and version the benchmark before using the revised rule to compare models.

B1. Review records are bound to the current candidate hash. A changed public row requires a fresh review, rather than inheriting a verdict about an earlier version.

Remove repeated incidents, preserve lineage

Structural checks enforced canonical labels, unique identifiers, complete three-level families, expected quotas, current independent reviews and required boundary metadata. Normalised exact matching and a conservative near-duplicate screen then identified repeated public content. Reviewer opinions alone were not blocking rules.

Original near-duplicate rule	Required condition
Normalisation	Unicode NFKC, case-folding and word extraction
Combined text length	At least 40 characters
Shorter / longer length	At least 0.93
Word-set Jaccard	At least 0.90
Request-only SequenceMatcher	At least 0.94
Combined request/context SequenceMatcher	At least 0.96

All members of flagged duplicate groups were removed, including every variant of an affected baseline family. Removing four complete families and fourteen boundary singletons left 1,974 examples. No replacement cases were generated merely to restore the round total of 2,000. The retained quotas and exclusion list became part of the benchmark identity.

The append-only authoring history is larger than the scored cohort: 2,087 records comprise 2,000 original revisions and 87 later variants. All final scored rows are original revisions. Later training-data generation excluded the entire history, including abandoned variants, rather than just the 1,974 evaluated examples. A changed row ID or lightly rewritten request does not make a scenario independent.

The original duplicate filter targets obvious near copies, not every semantic paraphrase. Later production added broader family screening. The historical benchmark was also inspected repeatedly during development, making it a development-used regression asset. The separately frozen 999-example test supplied fresh held-out evidence for the training study.

Preserve the unit that carries dependence

Family membership matters both before and after scoring. Keep related cases within one split, and retain whole-family dependence when estimating uncertainty. The original benchmark's 982 components are the resampling units used in Paper 2; treating all 1,974 rows as independent would overstate the effective scenario count.

B1, D1. The benchmark remains synthetic and English-only. It tests this routing policy, not the full distribution of live enterprise requests. Source/history files, exclusions and exact hashes are retained.

Recipes turn coverage into an explicit plan

The production recipe declared the task, category and split quotas, challenge mechanisms, teacher weights, reasoning levels, provider bindings, budgets and quality rules. The planned mix was 10% easy, 20% medium, 30% hard, 20% confusing and 20% conflicting. These are generation strata, not measured difficulty labels.

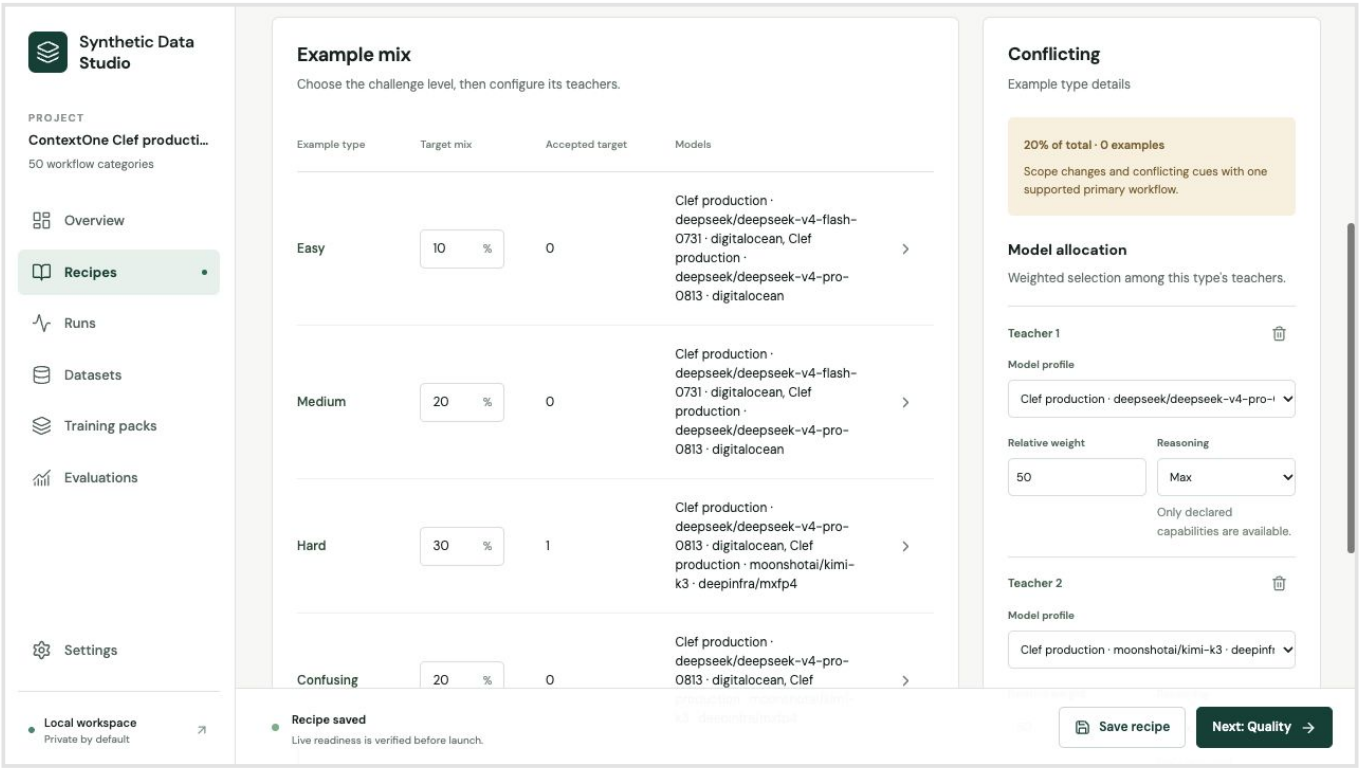


Figure 6. Studio's example-mix and model-allocation controls. This retained view is a two-example refill recipe using the campaign's mix and routes; its per-cell counts are not the final export. The frozen campaign membership, rather than the displayed refill total, determines study coverage.

Frozen split	Easy	Medium	Hard	Confusing	Conflicting
Training, 4,998	498	1,000	1,500	1,000	1,000
Validation, 500	50	100	150	100	100
Sealed test, 999	99	200	300	200	200

All fifty classes occur in each split. Training has 100 examples per class except categories 30 and 48, with 99 each. Validation has ten per class; test has twenty except category 48, with nineteen. The three-example shortfall was accepted rather than extending the campaign for negligible benefit.

D1, S1. Split seeds: 202611041 / 202611042 / 202611043. The planned 70% harder strata do not imply that 70% were empirically difficult; Paper 2 reports observed test performance.

Choose teachers and qualify the actual routes

DeepSeek Flash supplied inexpensive variation for simpler cases. Pro handled most authoring and the conditioned target pass. Kimi supplied a different model-family perspective in difficult authoring and every blind review. Routing was pinned through OpenRouter, so a nominal model name could not silently substitute an untested provider.

Production role	Exact model / serving route	Exported authors
Easier authoring	DeepSeek V4 Flash 0731 / DigitalOcean	1,173
Majority authoring; repairs; conditioned targets	DeepSeek V4 Pro 0813 / DigitalOcean	3,771
Difficult authoring; blind review	Kimi K3 / DeepInfra MXFP4	1,553

Stratum	Flash author share	Pro author share	Kimi author share
Easy	80%, low	20%, low	—
Medium	50%, high	50%, high	—
Hard	—	80%, high	20%, high
Confusing	—	60%, high	40%, high
Conflicting	—	50%, max	50%, max

The weighted plan implied 18% Flash, 58% Pro and 24% Kimi. Actual exported shares closely followed it. Pro used high reasoning for repairs and target annotation; Kimi used high for blind review. Output allowances were 8,192 tokens for Flash and 16,384 for Pro/Kimi, with three attempts, one repair and ten concurrent requests.

Qualification changed the routes before scale-up. DeepInfra overloads, regional CoreWeave failures and Modal overloads made nominal price insufficient. DigitalOcean Flash passed two checks; Pro passed three checks, followed by 207 successful trial requests. This supported the chosen workload, not a provider reliability guarantee. Early production rate limits remained in the record.

The trial also exposed content defects: weak context integration, customer/supplier role confusion and answer-revealing “only X/not Y” phrasing. Instructions were tightened before production. Testing a route should therefore inspect both delivery reliability and the examples it produces.

D1. DeepSeek route quantisation was not declared; do not infer it from the earlier DeepInfra benchmark configuration. Kimi’s recorded production route was MXFP4, not the benchmark’s BF16 route.

Distillation rights require a use-specific decision

Output ownership does not by itself settle whether an enterprise may train another model on those outputs. The relevant questions are which model is being developed, whether it competes with the provider, whether it will be distributed, and which publisher and serving-provider terms govern the requests.

OpenAI restricts competing-model development but expressly permits classification models that are not distributed or commercially offered to third parties. Anthropic restricts competing products and models; its guidance includes noncompeting specialist classifiers among permissible examples. These are **not blanket synthetic-data bans**, and a claim that a recent change universally prohibits this work would overstate the published terms. OpenAI agreement; Anthropic terms; classifier guidance.

The provider makes the initial enforcement decision and can suspend access under its agreement. A customer's belief that its work is "not distillation" does not prevent an account restriction; contractual dispute routes remain available. Where a use falls near that boundary, written clearance or a suitably licensed teacher is a more prudent basis for an enterprise dependency than an optimistic interpretation of the exception.

AI One sourced the training prompts, contexts and labels from eligible DeepSeek and Kimi routes. OpenAI and Anthropic had evaluation roles; OpenAI tooling also supported engineering and retain/quarantine audits without rewriting retained training text or reference labels. The choice of teachers reduced uncertainty about the intended research use while keeping those roles distinct.

Control	What it establishes
Distillable-text eligibility enforced	Route selection obeys OpenRouter's eligibility signal.
Model/provider identities pinned; fallback disabled	Requests remain on the qualified route.
Publisher and serving-provider terms reviewed	Eligibility metadata is not treated as a licence substitute.
Use scope recorded	Internal research approval is not client redistribution clearance.

The exact DeepSeek releases carried MIT licences; Kimi K3 carried a custom licence. Recorded approval covered internal, nondistributed research. It did not establish unrestricted rights to distribute the resulting dataset or trained artefacts to clients. OpenRouter describes its eligibility filter as best effort; underlying terms still apply.

Terms checked for this paper on 6 November 2026. OpenAI §§3.3(e), 17 and 8.2; Anthropic D.4, I.3. OpenRouter distillation controls and terms. Publisher and route licences are linked on p. 20. This is the study's operational rationale, not a universal legal conclusion.

Separate authoring, target estimates and review

The author returned a public request and optional context, an assigned primary category, a short justification, competing categories and a decisive cue. Those private fields supported provenance and inspection. They were not added to the student's input as hints.

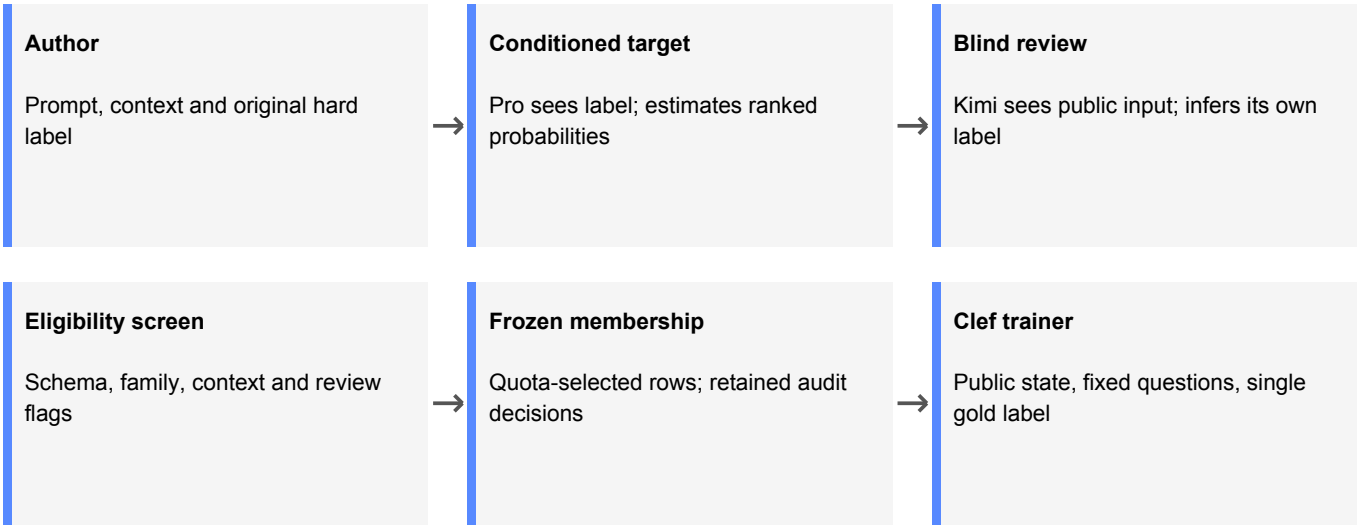


Figure 7. Executed production pipeline. The generic probability-target stage was paid for but its targets were not used in Clef's hard-label training.

The middle pass was conditioned on the author's label. Its ranked probability targets were part of Studio's generic recipe, not an independent confirmation of the assigned class. Clef's native training used a hard reference label, so those probability targets were discarded from the trainer export. Their generation cost nevertheless remains included in the measured spend.

Kimi's blind pass saw the public input and task, not the author label. That is a separate perspective, but it is not always a different author model: Kimi authored some difficult rows too. Benchmark cross-review, by contrast, always used the other author model. The two review designs should not be described as identical.

Requirement in production authoring	Intended effect
Contextual follow-up depends on its context	Avoid irrelevant context attached to an obvious request.
Integration uses at least two visible facts	Require information combination rather than extra length.
Boundary cue remains observable	Make one label defensible without private author intent.
Scope and stage have visible chronology	Separate completed work from the current request.
Quoted instructions are delimited	Treat embedded directions as data, not new authority.

D1. Ten mechanisms were exported: direct, indirect, contextual follow-up, workflow boundary, context integration, lexical decoy, deliverable versus topic, scope change, workflow stage and quoted instructions. Sectors varied; the campaign remained synthetic and English-only.

Curation is a recorded selection decision

The campaign completed 6,708 candidate records. Automatic eligibility checks covered the full population; a later detailed, OpenAI-assisted audit examined flagged cases. All completed rows had pipeline blind review, but it would be incorrect to claim that each received individual independent semantic adjudication or human approval.

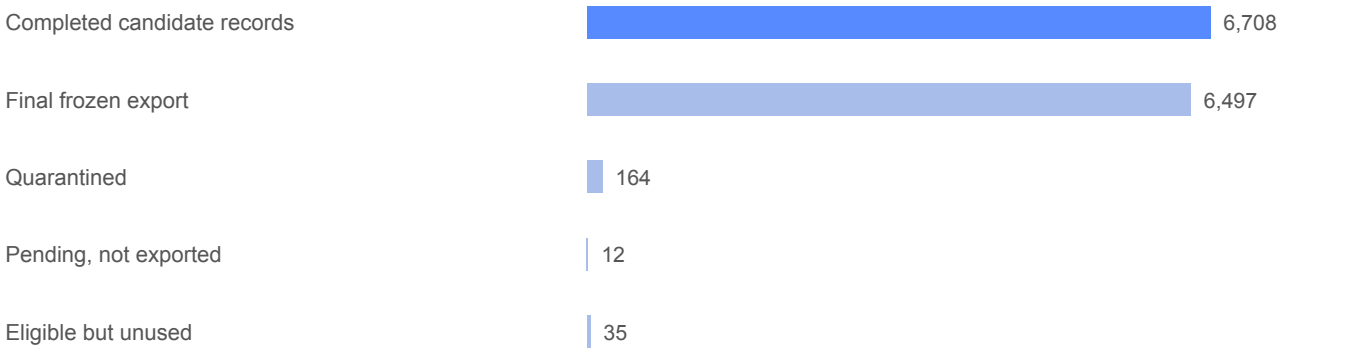


Figure 8. Candidate accounting: 6,708 = 6,497 exported + 164 quarantined + 12 pending + 35 eligible but unused. Bars share a zero baseline.

There were 6,193 automatically eligible candidates and 339 retained after detailed review, giving 6,532 eligible rows. Quota selection exported 6,169 from the first group and 328 from the second. The detailed audit recorded 503 decisions: 339 retain and 164 quarantine. Thirty-five otherwise eligible rows were not needed by the final quotas.

Forty-nine exported rows retained a blind-review category disagreement: 35 in training, four in validation and ten in test. The original author labels were preserved. Disagreement triggered attention, not a silent majority-vote rewrite or an automatic rejection rule that would remove all difficult cases.

The final family screen also compared candidates against 2,594 external exclusion entries. Sublinear-TF, smoothed-IDF cosine similarity, a 0.2 threshold and maximum postings of 100 located candidates for inspection. Similarity was a search aid; shared generic workflow language alone did not establish a reused incident. Retention decisions distinguished generic overlap from specific scenario reuse.

The earlier 222-row pilot illustrated why this matters. A deeper family audit found overlap despite no exact historical matches. It retained 211 rows, quarantined eleven and froze a 169/42 train/validation split. That was a feasibility pack with sparse class coverage, not the later production corpus; all 222 pilot rows were excluded from production.

D1. Automated audit decisions and explanations are preserved. Public text and retained author labels were not rewritten by that audit. The initial pilot creation cost is not established by its offline review.

Inspect examples without mixing their roles

Studio's dataset view puts the request beside its category, intended type and review state, with an inspector for prompt, context and separate evidence. This makes it possible to examine a suspicious boundary in its original form, rather than relying on a batch summary or a generated explanation detached from the source row.

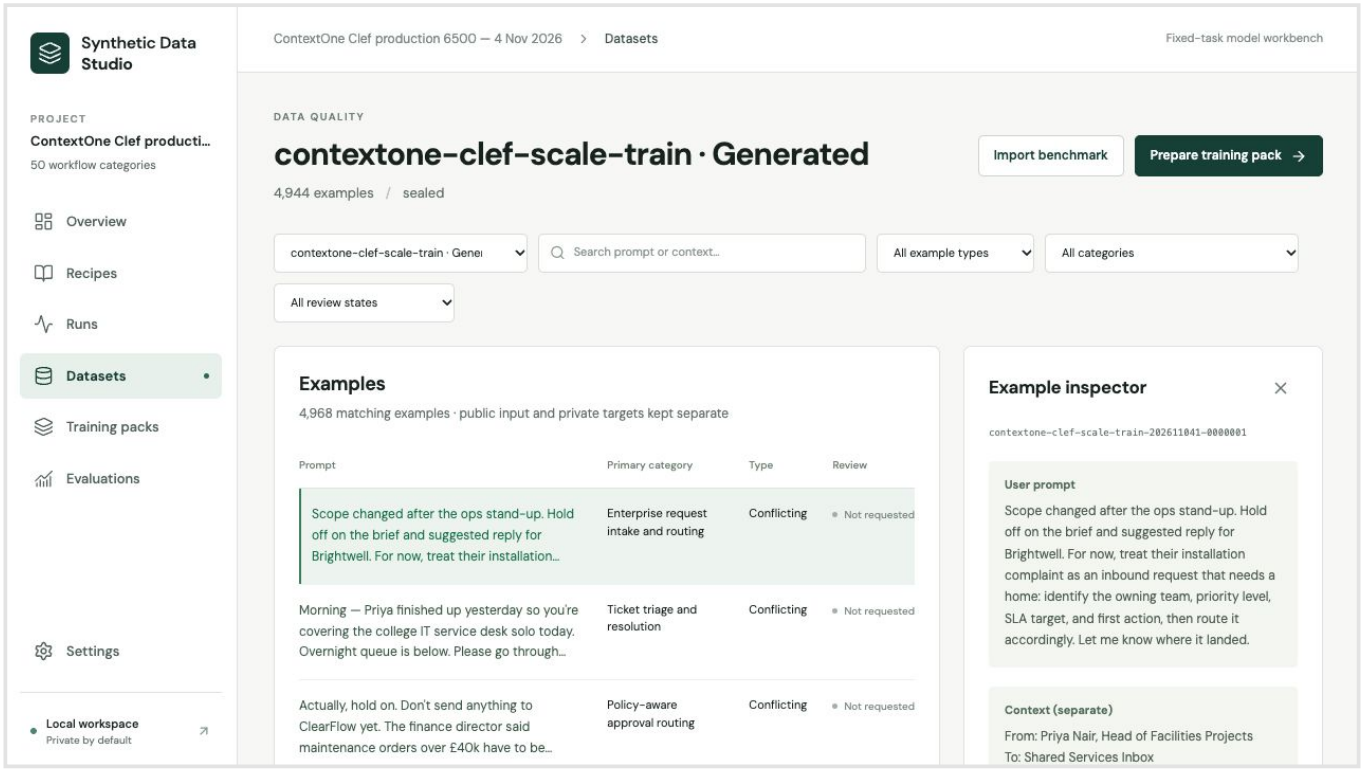


Figure 9. Retained main source dataset and example inspector, captured on 6 November. The header counts accepted source-run rows; the broader example list includes other candidate states. "Not requested" is the optional application review state, not proof that pipeline blind review was absent. This screenshot is not the final multi-run pack manifest.

This separation is useful even when most processing is automated. The public input can be compared with the private assigned category and reviewer finding without exporting those audit fields to the classifier. Review notes, provenance and raw evidence remain available for investigating why a row was retained or excluded.

The production programme combined the workbench with campaign-specific audit and freezing tools. The final decision receipts, membership manifest and source hashes are the authority for the reported counts. A green application status or a snapshot count does not substitute for those records.

For the enterprise operator, the useful review question is concrete: does the visible request support the recorded action under the declared policy? If not, quarantine the row or revise the policy in a new version. An elaborate author explanation cannot repair missing evidence in the public request.

D1, S1. Screenshots show actual application records. No review decisions, datasets or budgets were changed while preparing this paper. Optional human-review controls do not imply that human adjudication occurred in this study.

Freeze the handoff, not just the file format

A pack is the reproducibility boundary between data preparation and model training. It fixes membership, labels, task policy, encoder settings and provenance together. The production export combined fourteen source runs after curation; it was not simply a download of the main run's accepted examples.

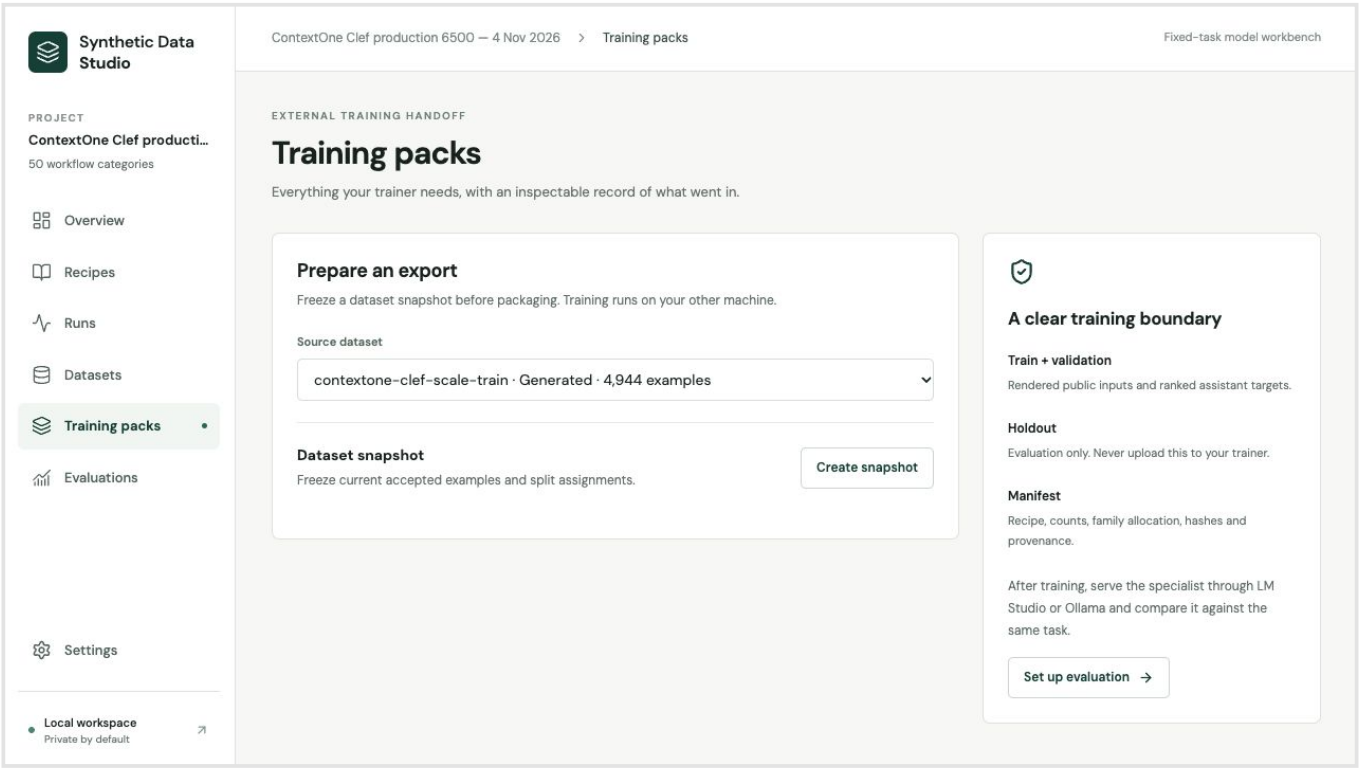


Figure 10. Studio's external-training handoff screen with the main source dataset selected. The generic interface describes ranked targets; the study-specific Clef export used the hard-label format on p. 17. No new snapshot was created for this paper. Final membership comes from the multi-run frozen manifest.

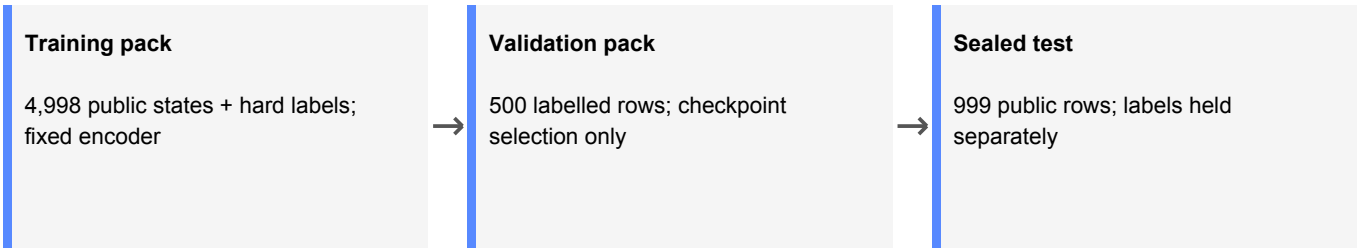


Figure 11. The export boundary. Public test inputs and private test references remain separate; training handoffs exclude holdout labels and audit explanations.

Selection used fixed category/bucket quotas and ascending eligible IDs per cell. A fixed order tag and example ID determined hash-based row ordering. Recipe, profile, template, membership and output hashes make the export reproducible.

Historical versions, pilot and trial candidates were excluded. Later splits also excluded preceding public versions, including rejected or repaired rows. Family and membership checks preceded release; training handoffs omitted test labels and private audit fields.

D1, S1. Final counts: 4,998 train / 500 validation / 999 sealed test. Only validation selected a checkpoint. The historical 1,974-example benchmark remained a separate transfer comparison, never a training pack.

Validate the input the model will actually receive

Short public prompts do not necessarily make short training sequences. Clef receives the policy and all fifty choice definitions as well as request and context. Token checks therefore used the complete rendered input and the pinned native encoder, rather than measuring the request text alone.

Split	Minimum	Median	95th percentile	Maximum	Total tokens
Train, 4,998	1,989	2,095.5	2,265	2,577	10,547,844
Validation, 500	1,993	2,097.5	2,248	2,577	1,055,462
Test, 999	1,991	2,100	2,254	2,510	2,110,544

Every example fit the 4,096-token limit. Truncation was rejected: cutting off a decisive context sentence or choice definition would change the task while leaving the label unchanged. Token and file-hash receipts bind the verified encoding to the exported bytes.

Clef trainer field	Content	Deliberately excluded
state	Public request and optional context	Private cue and author rationale
questions	Fixed policy and fifty named/described choices	Reviewer verdict and teacher identity
gold	Original single hard reference label	Teacher probability distribution

String choice identifiers followed the pinned encoder’s lexicographic order: 1, 10, …, 19, 2, 20, … . Replacing that with an intuitive numeric order would change the encoded decision problem. Encoding, category order and model revision were therefore treated as experimental settings, not interchangeable formatting choices.

The base was Cloudflare/clef-flash at revision fb4417d31c1d0f9e7e77f1141b40aaede7491e00. The vendored reference encoder was checked byte-for-byte against the pinned source. Native classification training retained Clef’s decision head; Paper 2 documents the adapters, optimiser, three epochs, validation selection and matched stock comparisons.

A good pack lets the trainer fail early on a missing field, hash mismatch or overlength row. It should not silently fill in a label, reinterpret the taxonomy or truncate data to keep a run moving. These checks are cheap compared with discovering an incompatible export after training.

D1, T1. Empty public context counts were 1,094 / 110 / 213: context was included where required, not padded mechanically. All three splits covered fifty classes.

Measure the cost of accepted evidence

The cost trial replaced token-budget guesses with observed usage. It completed 246 of 250 planned examples for $5.11165; all trial material was excluded from the production splits. Primary production then cost $156.93019, including repairs, failed work, refills and the probability-target stage unused by Clef.

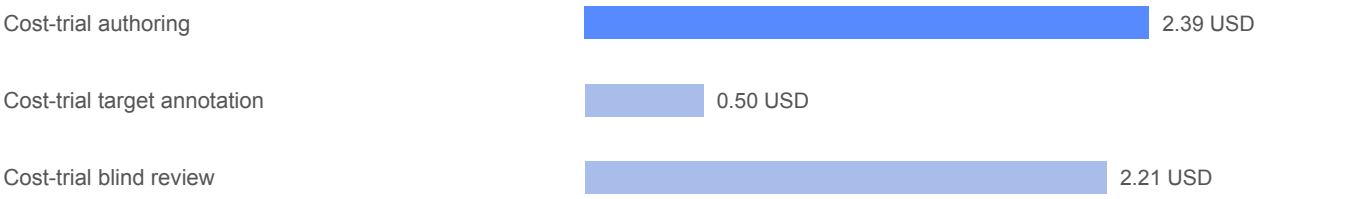


Figure 12. Measured trial-stage costs, rounded from the internal request ledger. The unrounded stage ledger differs from provider-reconciled trial spend by less than $0.001. This is not a production-stage allocation.

Cost measure	USD	Interpretation
Measured cost trial	5.11165	Provider account/key usage change
Primary production	156.93019	Provider-reconciled campaign total
Known trial + production	162.04183	Excludes original pilot creation, training and evaluation
Production per exported example	0.02415	Total production / 6,497 train, validation and test rows

Trial token evidence explains why estimates can miss badly. Cached tokens were 69.6% of 1,905,658 input tokens. Reasoning accounted for 81.4% of 626,670 output tokens. Reasoning is already included in output billing; adding it again would double-count. A provider’s headline rate without these workload characteristics is not a reliable accepted-example budget.

Production’s rounded response ledger total was $156.852490 over 20,189 settled attempts, $0.077698 below the reconciled account-level total used here. A further 508 unresolved attempt reservations totalled $42.727062. These were preserved for retry safety and **not added as confirmed extra spend**. An uncertain response must not be assumed free or automatically replayed.

Studio’s run records distinguish settled, in-flight and uncertain amounts, while the final campaign reconciliation establishes the paper’s monetary total. A full project budget also needs engineering, audit effort, training, evaluation and future serving costs. Paper 2 retains the $15,000 workstation and enterprise H200 scenarios; this page isolates data creation.

D1. Historical route prices and recorded usage, not a live quote. Figures exclude credit-purchase fees and tax. Original 222-example creation spend is unresolved; $162.04 is not an all-project total.

The ContextOne sample taxonomy

These are the fifty category names shared by the benchmark and new training corpus. Each ID also has a fixed definition in the versioned taxonomy. They identify the operation requested, not the user's industry. Domain-specific categories outside IDs 1–50 were excluded from this study.

IDs 1–25	IDs 26–50
1. Enterprise request intake and routing	26. Customer onboarding workflow
2. Ticket triage and resolution	27. Customer health and churn intervention
3. Policy-aware approval routing	28. Proactive incident communication
4. SOP execution agent	29. Multilingual customer workflow
5. Knowledge Q&A with citations	30. Voice-of-customer mining
6. Research brief generation	31. Lead enrichment and scoring
7. Meeting-to-actions workflow	32. Account research and call prep
8. Email/thread action workflow	33. Sales outreach sequencing
9. Document drafting workflow	34. RFP/RFI response workflow
10. Document redlining and comparison	35. Proposal generation workflow
11. Document extraction to structured data	36. CRM hygiene workflow
12. Form completion and submission	37. Quote and CPQ support
13. System-of-record update workflow	38. Discount and pricing approval
14. Data reconciliation workflow	39. Renewal and upsell trigger workflow
15. Master data maintenance	40. Campaign planning and content workflow
16. Exception investigation workflow	41. Personalisation and segmentation workflow
17. Escalation and handoff workflow	42. Competitive intelligence monitoring
18. Status reporting workflow	43. E-commerce recommendation workflow
19. Audit trail generation	44. Marketing compliance review
20. Workflow monitoring and recovery	45. Code change implementation
21. Customer support self-resolution	46. Code review workflow
22. Technical support troubleshooting	47. Test generation and maintenance
23. Customer case prioritisation	48. Build and CI/CD failure diagnosis
24. Refund, return and credit handling	49. Production incident triage
25. Complaint handling and de-escalation	50. Runbook execution workflow

B1, D1. The Multilingual customer workflow category is present, but the benchmark and production requests in this study are English-language. Category coverage is not evidence of multilingual performance. The frozen definitions and policy, rather than names alone, govern the classification.

Reproducibility and the research series

B1 — Original benchmark. TypeSafe.ai BENCHMARK_DESIGN.md, executable pipeline/boundary/dataset modules and the frozen pilot-gpt6-luna-20260922 run. Retained evidence includes taxonomy, prompts, inputs, labels, candidate history, reviews, approval and duplicate exclusions. Lineage: 16 September baseline → 17 September boundaries → 18 September pruning → 22 September GPT-6 comparison. The later comparison reused the same public inputs and references.

D1 — Data creation. Synthetic Data Studio's contextone-clef-v1 pilot audit/package and contextone-clef-v2 recipe, cost-trial and production notes; contextone-clef-production-20261104 completion, billing, final audit and frozen-pack manifest. The evidence register is docs/experiments/clef-followon-paper-evidence.md. Primary creation/pilot records are preserved in a 560-file archive with a checksum receipt.

S1 — Application and screenshots. Studio README, docs/DECISIONS.md, docs/RUNNING.md and read-only local application captures, 6 November 2026. Screens show retained source records and generic workbench controls; the final study packs also used campaign tooling. **T1 — Training contract.** Clef model card, pinned encoder and production training receipts. Training outcomes and compute economics appear in Paper 2.

Byte-level identity	SHA-256
Benchmark inputs.jsonl	091c30f511414b48e031d1514b132a441180e5e37e9794f8b7d8a0b0f7d26fb4
Benchmark labels.jsonl	1d13f413d97d73336c74a6525694c653658defe64a5b68b0065ed91d85d0c0b5
Benchmark taxonomy.json	f8525041303ea98e07c7bbe2df27e62974b4b3f4512632ef3c8ee9f1655b5927
Production pack manifest	2db1c657540c96a24312027e68528cd8c34a474c1f2dd2663d70a7e7dd310702

Publisher and route terms. DeepSeek Flash licence, DeepSeek Pro licence, Kimi K3 licence, DigitalOcean service terms and DeepInfra terms. Frontier and OpenRouter terms are linked on p. 12. Rights assessments are specific to the recorded use and date.

Three papers, three contributions. Paper 1 evaluates stock decision models and query design. Paper 2 measures the uplift from task-specific Clef training and its enterprise economics. This paper explains how the benchmark and separate training corpus were constructed, audited and frozen. Together they connect the result to its evidence without treating a synthetic reference set as universal enterprise ground truth.

Author: Mark J Sykes, AI One Labs. Prepared 6 November 2026. Underlying research records remain archived; publication of this account does not itself grant redistribution rights over datasets, model outputs or trained artefacts.