OMNICENS: SELF-IMPROVING DISTILLATION FOR MULTIMODAL CLASSIFICATION OF MILLIONS OF PROFILES

Anonymous authors
Paper under double-blind review

ABSTRACT

Classifying millions of social-media profiles from images, text, and contextual evidence is valuable for audience measurement but expensive when every prediction depends on a frontier model. We propose OmniCens, a self-improving distillation framework for reducing this dependence through local open-source multimodal models. Its central principle is that an expensive teacher response can serve both the current classification and subsequent student learning. We formulate the objective as minimizing amortized classification cost under precision, recall, and coverage constraints, accounting for teacher inference, calibration, student updates, and local execution. A confidence-based acquisition rule directs teacher supervision toward uncertain profiles, and repeated distillation targets recurring weaknesses while retaining the complete input evidence and attribute schema. The evaluation protocol isolates the value of continued learning and selective supervision through comparisons with fixed distillation, non-learning cascades, and random acquisition under matched budgets. Independent annotations assess predictive quality separately from teacher agreement, while held-out profiles and communities assess generalization. The key question is when reductions in teacher dependence recover the cost of adaptation without degrading classification quality. OmniCens connects this trade-off to a practical objective: making detailed audience characterization and creator–brand matching feasible for researchers and organizations with limited computational resources.

1 INTRODUCTION

A social-media profile is not a single image or document. Its evidence is distributed across photographs, captions, descriptions, metadata, and engagement records. Audience measurement adds two complications: the same person can appear in several creators' audiences, and the requested output can contain many categorical fields with different admissible values. An efficient classifier must preserve both the evidence and the task. Dropping difficult fields or reading fewer images can reduce cost while changing what is being measured.

A frontier vision–language model can process this heterogeneous evidence, but calling it for every profile makes cost grow directly with the population. A local student changes that calculation. A teacher response can answer the current query and become a training example for later queries. The resulting system is worthwhile only if the avoided teacher calls repay preparation, training, calibration, and local inference costs at the required level of quality.

OmniCens organizes this process around complete profile snapshots, reusable teacher supervision, and a structured local predictor. Each snapshot records the available images and text, the field vocabulary, and the person's creator associations. Predictions are frozen before the corresponding teacher query. Subsequent updates can therefore be compared with the student that existed before those labels were acquired.

The empirical question is whether adaptation adds useful predictive information beyond class frequencies. Our pilot makes that distinction visible. In the first prospective round, the source-trained model agrees with 60.99% of known teacher decisions, compared with 58.73% for the current-training majority. Adaptation reduces agreement to 56.67%. In the next round, continued adaptation

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

improves on the held-fixed student by 0.50 percentage points but remains below the updated majority. These results identify a concrete obstacle to the proposed cost reduction: a cheaper predictor has not yet earned the right to replace the teacher.

Contributions. The paper provides (i) an implemented workflow that preserves complete stored evidence and a fixed output schema across acquisition and local learning; (ii) a reconstruction of archival supervision that recovers 187,207 explicit negative or not-applicable targets; and (iii) two prospective prediction rounds that compare adapted students with held-fixed students and class-prior controls while measuring coverage. The resulting system supplies the inputs, saved predictions, and accounting needed to test selective distillation without obtaining an apparent saving by reducing the task.

Scope. The title identifies the intended deployment scale. The registered abstract describes the full proposed evaluation protocol. This paper reports the implemented components and a pilot, not a million-profile experiment. Calibrated uncertainty routing, matched-budget acquisition comparisons, independent annotations, and held-out-community evaluation remain uncompleted. We distinguish the proposed deployment objective from the evidence for it throughout.

2 Related work

Knowledge distillation transfers behavior from a larger predictor to a smaller one (Hinton et al., 2015). OmniCens uses teacher-generated categorical labels rather than teacher logits. Cost-aware cascades route queries among models with different capabilities and prices (Chen et al., 2023). More directly, Online Cascade Learning updates a smaller model from large-model demonstrations while learning when to defer (Nie et al., 2024). Cache & Distil continuously trains local students as neural caches and uses active learning to select API queries for classification (Ramírez et al., 2024). Reusing a purchased response for subsequent learning is therefore established prior work. Our study examines the evidence-preservation and evaluation problems that arise when each query is a complete multimodal profile with a large structured output. It does not establish an advantage over either online method.

Selective prediction makes the coverage–risk trade-off explicit (Geifman & El-Yaniv, 2019), while calibration research shows why confidence alone cannot certify correctness (Guo et al., 2017). Language models can supply useful annotations on particular tasks (Gilardi et al., 2023); that finding does not establish the accuracy of a multimodal profile teacher. We consequently separate agreement with the teacher from accuracy against independent references. Multiple-instance learning provides a further connection: a bag of observations is aggregated into a prediction (Ilse et al., 2018). Here the observations are image and text instances and the output has field-specific admissible values. The pilot's contribution is the complete-profile workflow, supervision reconstruction, and prospective controls, rather than a new distillation or aggregation primitive.

3 Problem and accounting

Let $x_i = (\mathcal{I}_i, \mathcal{T}_i, m_i)$ denote a deduplicated profile, containing all retained images, all retained text segments, and contextual metadata. Let G_i be the set of associated creators. The application schema has 1,000 categorical positions: 961 eligible fields and 39 structurally excluded positions. Across the eligible vocabularies there are 4,762 field–value candidates. Formally, the schema is a collection of categorical fields $j \in \mathcal{J}$, with vocabularies $\mathcal{V}_j$. An output is $y_{ij} \in \mathcal{V}_j \cup \{\text{UNKNOWN}\}$. A fixed exclusion mask forces disallowed sensitive-field outputs to UNKNOWN. Excluded positions remain present as UNKNOWN so that an output cannot silently change the task.

The teacher T produces weak labels $\tilde{y}_i$. The student S_{θ_t} produces field probabilities and a coverage decision. A profile is accepted locally only under a declared routing rule; otherwise it is referred to the teacher. Let $a_{it} = 1$ denote a teacher query. The cumulative expense through N profiles is

$$\mathcal{C}_N = \mathcal{C}_{\text{setup}} + \sum_t (\mathcal{C}_{\text{update},t} + \mathcal{C}_{\text{calibration},t}) + \sum_{i=1}^{N} (\mathcal{C}_{\text{prepare},i} + \mathcal{C}_{\text{student},i} + a_i \mathcal{C}_{\text{teacher},i}) . \quad (1)$$

All terms in this scalar objective require the same declared unit, such as monetary expense under a fixed price schedule. The pilot reports counted tokens, CPU time, wall time, and cash outlay separately; it does not add these quantities or claim a measured combined scalar saving. Retries, incomplete responses, duplicated context, and unsuccessful training runs remain in this accounting. A preprocessing cache avoids repeating computation only when the complete content and preprocessing specification match. It does not turn a new evidence snapshot into an already-classified profile.

We seek to minimize $\mathcal{C}_N/N$ subject to field-wise or declared aggregate constraints on precision, recall, and coverage against independent labels. Such labels are not available in the pilot, so the implemented diagnostics measure agreement with $\tilde{y}$, not the constraints themselves. A model that emits only UNKNOWN can avoid false assertions while offering no useful coverage; it is therefore included as a control, not a successful solution.

3.1 Break-even condition

For interpretation, suppose a common per-profile teacher cost is c_T, a student cost is c_S, and adaptation plus calibration costs U. If adaptation reduces the teacher-query rate from q_0 to q_1 over M subsequent profiles while leaving the other expenses unchanged, the resulting cost change is

$$\Delta\mathcal{C} = U - M(q_0 - q_1)c_T. \tag{2}$$

Thus an update pays for itself only if $M(q_0 - q_1)c_T > U$. This is an accounting identity under the stated assumptions, not a measured improvement or a quality guarantee. With normalized local and preparation cost λ, amortized adaptation cost u, and teacher-query fraction q, a tenfold saving relative to teacher-only inference requires $\lambda+u+q \leq 0.1$. It is impossible if the local and adaptation terms alone exceed this budget. Quality constraints must also hold; the inequality cannot establish them.

4 OmniCens workflow

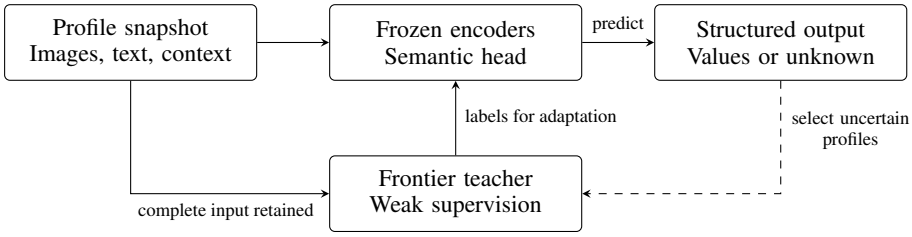


Figure 1: OmniCens operating loop. Solid paths correspond to implemented input, prediction, teacher-labeling, and adaptation components. Dashed uncertainty routing is part of the proposed deployment protocol; a calibrated router and a complete selective-learning cost advantage are not established by this pilot. Teacher labels remain weak supervision.

4.1 Complete evidence and cohort construction

The cohort is constructed from liker–follower records associated with a selected set of 2,902 creators. Membership includes the project's recorded relationship types; a repeat-liker proxy is not silently upgraded to a verified follow. The individuals being classified are audience members, not the creators. A person is deduplicated across associations, while G_i is retained. Scheduling favors coverage of additional creators rather than exhausting a small number of audiences.

A frozen metadata audit of 214 source-complete eligible training records finds 236 person–creator associations spanning 236 creators. These are prepared-source counts, not completed classifications. If the evidence snapshot, schema, and method are identical across associations, one person-level prediction replaces repeated association-level queries: 214 instead of 236 calls, a conditional 9.32% reduction. The campaign already deduplicates; this is a counterfactual call-count comparison, not an additional measured token, time, or quality gain over a unique-person baseline.

An eligible record for new rolling acquisition contains more than ten distinct readable images belonging to its declared evidence snapshot. Every available caption, description, comment, metadata item, and contextual text segment in that snapshot is retained. "Complete" refers to the stored snapshot, not to all information that exists on the platform. The evaluated snapshots contain images and text; complete video and audio streams are not evaluated. An unavailable item is recorded as unavailable rather than invented or counted as observed. Images are resized with aspect ratio and full frame retained, with a maximum dimension of 512 pixels. The originals are retained separately. This transformation is fixed across the compared conditions; quality parity with higher-resolution input is not assumed.

4.2 Teacher supervision and output validity

Each teacher packet contains the full applicable classification rules, field vocabularies, profile evidence, and individual image inputs. The pilot uses an authorized native interface to Grok 4.7: Extra High in the early learning stream and High Fast in the later delivery batches. The observed model labels, usage-export identifiers, and displayed input-read operations are recorded. These records do not expose hidden routing or the provider payload. A visible image-read record supports interface-level delivery; it does not prove model attention or unmodified retention through automatic context summarization.

We therefore distinguish complete fresh-input observations from engineering observations in which all inputs were delivered but context retention is unverified. Cases missing unique evidence are excluded from complete-input comparisons. A result also undergoes strict validation of subject identity, field positions, vocabulary, exclusion mask, and explicit unknowns. Validation establishes structural correctness only. The predictive evaluation uses the resulting labels separately from these structural checks.

4.3 Student learning

The local learner consumes frozen image and text representations of the complete stored input. The pilot uses CLIP ViT-B/32 image features (Radford et al., 2021) and BGE-small-en-v1.5 text features (Xiao et al., 2024). Each image is retained individually; all text is divided into contiguous chunks rather than truncated. The image encoder's positional embeddings are interpolated for a padded 512-pixel input, an experimental adaptation of its 224-pixel training resolution. Label support is estimated from training observations only. Explicit admissible negative or not-applicable labels remain training targets; UNKNOWN and unavailable targets are censored. This distinction matters when a weak teacher produces sparse outputs.

For field j, a training mask m_{ij} selects known admissible labels and w_{ij} is the recorded reliability weight. The corrected objective (the earlier reported head used inverse-class-balanced cross-entropy) averages a reliability-weighted categorical loss over active fields:

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{J}_+|} \sum_{j \in \mathcal{J}_+} \frac{\sum_i m_{ij} w_{ij} [-\log p_\theta(\tilde{y}_{ij} \mid x_i, j)]}{\sum_i m_{ij} w_{ij}} + \frac{\lambda_2}{2} \|\theta\|_2^2. \tag{3}$$

Here $\mathcal{J}_+ = \{j : \sum_i m_{ij} w_{ij} > 0\}$ contains fields with positive total supervision weight. Missing supervision is never replaced by a negative class. A prior-aware variant adds a fixed log class prior estimated from the training portion to a low-rank learned semantic residual. For prior-aware heads, classes unseen in training are excluded from normalization and prediction; a field with no supported class returns UNKNOWN. The support-≥ 2 rule evaluated below is an additional selective filter. The frozen feature encoders are part of the inference system; the small trainable head is not the full model footprint.

4.4 Acquisition, adaptation, and prequential evaluation

The proposed acquisition rule prioritizes uncertain profiles subject to budget and creator-coverage constraints. The proposed evaluation specifies three comparisons: random acquisition to test the value of selection, fixed distillation to test continued learning after the initial seed, and a non-learning cascade to separate adaptation from routing. Their purpose is to identify which component reduces

Table 1: Four-profile prequential development after two training profiles. Agreement and coverage use 3,611 known teacher decisions; selective agreement uses each row's answered-known subset; nonmajority agreement uses 1,121 decisions differing from the training majority. These percentages are not human-accuracy estimates.

Predictor	Agreement	Coverage	Selective	Nonmajority
Training majority	68.24	99.28	68.73	0.00
Shared semantic	51.87	100.00	51.87	18.82
Shared semantic, support ≥ 2	38.38	46.39	82.75	0.00
Shuffled semantic	21.74	100.00	21.74	21.50
Shuffled semantic, support ≥ 2	10.99	14.57	75.48	0.00
Always unknown	0.00	0.00	–	0.00

total expense at matched quality and resource budgets. The pilot below does not execute these three acquisition arms.

Pilot acquisition uses breadth-oriented scheduling; uncertainty-based routing has not yet been evaluated. The prospective protocol freezes student predictions, class support, model state, and decision rule before the corresponding teacher query. That freeze precedes the teacher query for each of the ten new profiles reported in two successive rounds below; it does not describe every acquisition batch. The second round evaluates continued learning after the first round is scored, with the earlier student retained as a fixed comparator. Teacher answers are released subsequently for evaluation and later training. A complete quality evaluation requires calibration and evaluation profiles disjoint from training. Profiles linked across multiple creators require deduplication before splitting; a held-out-community evaluation additionally separates creator-associated groups. The present small development comparisons do not substitute for that evaluation. Previously assigned calibration and test roles are unused in this pilot; the four earlier development-evaluation profiles subsequently join the six-profile adaptation set.

5 Experiments

5.1 Evaluation units and reference quality

All predictive results below use weak teacher labels. The early current-method stream uses Grok 4.7 Extra High; the later delivery batches use High Fast. These conditions are not pooled as one quality benchmark. The first development experiment fits on two profiles and records predictions on four additional profiles before the evaluation worker decodes their references. Other project workflows had already accessed some references, so this is prequential development rather than an untouched test set. One of the four development profiles has exactly ten images and does not meet the later rolling-acquisition criterion of more than ten; this historical exception is retained and disclosed. The four profiles contain 3,611 known reference decisions, 43 explicit unknowns, and 190 unavailable decisions from one unavailable output block. The statistical sample size is four profiles, not 3,611 independent subjects.

5.2 Development comparison

Table 1 compares the shared semantic head with a shuffled-semantic control and a constant training-majority predictor. Known-reference agreement counts abstentions as misses. Coverage is the fraction of known reference cells receiving a prediction; selective agreement conditions on those answered cells. "Support $\geq$ 2" abstains unless the predicted class has at least two known training observations. It does not replace an unsupported winner by another class.

Semantic alignment raises raw agreement from 785/3,611 to 1,873/3,611 relative to shuffled semantics. However, the constant majority achieves 2,464/3,611. The support-filtered semantic model agrees on 1,386 of 1,675 answered known cells (82.75%) while leaving over half of the known reference cells unanswered. This is a coverage trade-off, not preserved full-task quality. Among 1,121 nonmajority reference cells, the raw semantic model agrees on 211 and the shuffled control on 241;

the aggregate benefit therefore does not establish superiority on the minority decisions. The two raw heads answer all 43 explicit teacher-unknown cells, while the two support-filtered heads answer none. A disagreement with teacher unknown is not automatically a factual error.

5.3 Reconstructing archival supervision

An exploratory comparison applies an archival-source head after 40 updates to the six current development profiles. That head agrees with 1,328/5,502 known teacher decisions (24.14%); the earlier current-six head agrees on 2,404/5,502 (43.69%), and the six-profile majority on 4,229/5,502 (76.86%). These are development diagnostics with unequal training histories, not a matched transfer experiment. No field/ordinal/vector mapping discrepancy was found in the corresponding audit.

The audit instead identified a mismatch between target conditions: the earlier declared source recipe censored explicit admissible negative and not-applicable values, whereas the current task treats them as known answers. Reconstructing the archived supervision from original records recovers 187,207 explicit targets while preserving 581,969 existing positives. The corrected 822-profile source has 769,176 known targets, 410 explicit unknowns, and 20,356 unavailable cells. Its cached evidence contains 4,025 images and 9,898 text chunks comprising 4,836,393 tokenizer tokens. This historical set does not uniformly satisfy the newer more-than-ten-image acquisition criterion. Recovery does not impute missing values. The 822 profiles are a separate historical transfer set; membership in the current liker–follower population is not established, and creator-community overlap is not excluded. An existing metadata audit finds no exact recorded-person, owned-post, or file-hash collisions between these 822 profiles and the 493-person rolling pool containing the prospective profiles. This does not certify real-world entity or creator-community disjointness. The corrected prior-aware condition is evaluated separately from the preserved failed model.

5.4 Teacher execution cost

A completed native delivery batch produces structurally valid outputs for 18 profiles over 2,338 observed wall seconds and consumes 22,009,953 counted tokens. Four profiles have audited fresh-input outputs released to the primary current-method set; four additional outputs are released only in explicitly separate engineering strata. Other outputs remain outside that primary release. A subsequent uninterrupted batch produces 16 structurally valid outputs over 1,289 observed wall seconds and consumes 14,045,083 counted tokens. Its complete-input and output-provenance audit is separate from structural validation. Neither batch's structural output count is an accuracy score, and their different conditions do not identify a causal speedup.

The second batch corresponds to approximately 44.7 structurally valid profile outputs per observed hour in this specific native workflow. Both totals include repeated and cached context. Usage-export timestamps and model identifiers support conservative whole-turn accounting but do not expose provider-side chat identity or per-profile token attribution. No costs are discarded because a response is incomplete or unusable.

5.5 Local computation and prospective evaluation

The shared head contains 14,344 trainable parameters, or 114,752 bytes of float64 weights before metadata. This excludes the frozen encoders and semantic cache. On two new profiles, extracting all 32 images and 111 text chunks (55,972 content tokens) takes 30.306 process-tree CPU seconds and 151.983 wall seconds at an observed 19.94% of one CPU core, with 1.187 GB peak resident memory and normal observed thermal state. No GPU is used in this measurement. The recorded interval covers representation extraction; prediction and preparation are separate costs.

The corrected source condition uses a predeclared 40-update snapshot and identical 20-update warm and cold adaptation on the six current training profiles. Source priors are not carried into the current adaptation: both arms recompute priors and class support from the same current supervision. Predictions on new profiles are frozen before teacher-label release. This prospective comparison evaluates the corrected condition; it does not isolate the effect of target reconstruction from the changed priors and objective.

Table 2: Six prospective profiles with the same frozen models. Known agreement counts abstentions as misses; coverage conditions on 487 known references. Answers on 5,279 teacher-unknown cells are reported separately.

Predictor	Known correct	Agreement (%)	Coverage (%)	Unknown answered
Source reliability prior	292/487	59.96	100.00	5,279/5,279
Source residual, update 40	297/487	60.99	100.00	5,279/5,279
Current-six majority	286/487	58.73	100.00	5,267/5,279
Cold prior-aware adaptation	278/487	57.08	100.00	5,267/5,279
Warm prior-aware adaptation	276/487	56.67	100.00	5,267/5,279
Cold, support ≥ 2	275/487	56.47	95.48	5,229/5,279
Warm, support ≥ 2	273/487	56.06	95.48	5,229/5,279
Preserved shared head	85/487	17.45	100.00	5,279/5,279
Always unknown	0/487	0.00	0.00	0/5,279

5.6 First prospective round

Six new profiles were evaluated in three successive pairs using identical frozen models: each pair's predictions preceded its teacher dispatch, with no model update between pairs. All three references use Grok 4.7 High Fast and observed fresh delivery of all current source text and images. Known decisions per pair are 183, 159, and 145, respectively. Combined teacher-known coverage is 487/5,766 (8.45%), with 5,279 explicit unknowns and no unavailable cells. The 39 structurally excluded fields per profile are outside these denominators. These results remain separate from the earlier Extra High development stream.

Warm versus cold known-correct counts are 100 versus 102, 85 versus 85, and 91 versus 91 across the three pairs. In aggregate, warm adaptation agrees on two fewer known decisions, a difference of $-2/487 = -0.41$ percentage points; both adapted models trail the current-six majority control. The source residual exceeds its person-invariant prior by 5/487 decisions (1.03 percentage points), an insufficient basis for a reliable benefit claim with six people. At support ≥ 2, both adapted heads answer 465/487 known decisions (95.48%); selective agreement is 273/465 (58.71%) for warm and 275/465 (59.14%) for cold. The experimental unit remains the person: six observations support these paired comparisons.

The high answer rate on teacher-unknown cells shows that the current support rule does not reproduce this teacher's abstention behavior. Such disagreement does not establish that student answers are factually false. Conversely, always predicting unknown matches 5,279/5,766 cells (91.55%) while answering no known decision, illustrating why overall agreement including unknowns is insufficient. Teacher-known coverage, known agreement, selective coverage, and unknown-answer behavior must be reported together. The observed support rule is therefore insufficient as a local acceptance policy.

A bounded audit of the first four references found the original decision-rule prefix unchanged and ordinary neutral values still present in the outputs; it did not establish a concrete semantic violation. Nevertheless, the global UNKNOWN extension and sparse-output complement leave the boundary between an ordinary neutral value and abstention incompletely specified. The much lower known coverage than in the earlier Extra High stream may reflect effort, serialization, context, or profile differences; these factors were not isolated. Here, strict refers to observed source delivery and schema validity, not certified semantic-policy conformance or human correctness. These comparisons measure agreement with this particular teacher condition, not factual student failure or preservation of the original method's quality.

5.7 Second prospective round

After scoring the first six prospective profiles, we appended their known teacher decisions to the original six-person TRAIN set. A declared continuation used exactly 20 additional updates, fresh Adam state, and recomputed priors and support on all twelve profiles; the regularization and architecture were unchanged. This changes both the residual and the priors, so it does not isolate the contribution of either component. The update took 27.15 seconds of wall time and 5.319 process-

Table 3: Second prospective round: four people, identical future inputs, and frozen predictions. Counts concern weak-teacher agreement.

Predictor	Known correct	Agreement (%)	Unknown answered
Held-fixed warm six	1,212/1,816	66.74	2,020/2,028
Warm continuation, twelve	1,221/1,816	67.24	2,020/2,028
Current-six majority	1,255/1,816	69.11	2,020/2,028
Current-twelve majority	1,271/1,816	69.99	2,020/2,028
Source residual, update 40	1,277/1,816	70.32	2,028/2,028
Fixed zero-shot cosine	175/1,816	9.64	2,028/2,028

tree CPU seconds under the declared local duty limit. Predictions from this continuation and all earlier controls were frozen before the next four people's teacher dispatch.

These four fresh-window references contain 1,816 known decisions out of 3,844 eligible cells (47.24%), 2,028 explicit unknowns, and no unavailable cells. They are a separate prediction round, not pooled with the first six or with references whose delivery crossed a summary. All raw predictors in Table 3 answer all 1,816 known decisions.

The twelve-person continuation gains 9/1,816 known decisions over the held-fixed six-person warm model, or 0.50 percentage points, while remaining below the current-twelve majority control. Paired changes for the four individual people are 0.00, -0.16, $+0.73$, and $+3.03$ percentage points; their unweighted mean is $+0.90$ points. With four paired observations, the aggregate change is descriptive. Both warm models answer 2,020/2,028 teacher-unknown cells, so this continuation does not resolve abstention disagreement. The change in teacher-known coverage from the preceding round further limits comparisons across rounds. The interpretation remains agreement with the recorded teacher condition.

6 Limitations and deployment implications

The main limitation is empirical: the prospective evaluation contains ten people in two rounds, and its references are model-generated labels. Field decisions within a person are correlated. The thousands of categorical decisions therefore do not supply thousands of independent observations. Teacher agreement measures imitation under the recorded teacher condition; it cannot establish factual accuracy, fairness, or safe replacement of the teacher.

The experiments also leave the central efficiency question open. Acquisition follows creator coverage rather than a calibrated uncertainty policy. The local head is small, but feature extraction and preparation remain part of its cost. Neither the source model's modest advantage over a majority control nor the continued student's gain over its predecessor establishes quality-preserving savings. A release decision needs an acceptance rule chosen on separate calibration data and tested against independent labels.

Teacher execution presents a further reproducibility constraint. The native interface can summarize context, and its visible tool history does not expose the provider's final payload or internal model state. We retain separate evidence strata for fresh-input execution and complete delivery with uncertain retention. Counted usage includes cached and repeated context; it is not equivalent to uncached tokens or dollar cost. The serial interface measurements also cannot predict API-batch throughput. Prepayment changes the immediate cash charge, not the amount of computation consumed.

Finally, the profile data and field vocabulary are restricted. This prevents independent reproduction of the application-specific experiments from the anonymous supplement alone. The complete-input claim concerns the stored snapshot, not all platform activity or a person's offline characteristics. The pilot does not support population-wide demographic or sensitive-attribute inference.

7 Conclusion

OmniCens treats a teacher query as both an immediate classification and a possible investment in later local predictions. Its accounting makes the condition for that investment explicit: avoided

teacher expense must exceed adaptation and execution costs while quality and coverage remain acceptable. The implemented pilot preserves complete stored evidence, reconstructs missing negative supervision, and evaluates successive students prospectively.

The present students do not meet the replacement criterion. Continued adaptation gains nine known decisions over its predecessor but remains below the majority control. The source-trained predictor is the strongest learned model in both prospective rounds, and its advantage over simple priors is small. These findings focus the next evaluation on calibrated local acceptance and useful departures from class-frequency predictions. They also show why complete-input accounting, prior baselines, and abstention measurements belong in an efficiency claim from the outset.

AI USE STATEMENT

Generative AI assistants contributed to research ideation, experimental design, implementation, debugging, data-format audits, interpretation, literature retrieval, and drafting and formatting the manuscript. Grok supplied the weak labels. AI assistants also helped develop and check the mathematical expressions and tests. Automated checks covered saved predictions, schema constraints, numerical calculations, source correspondence, and document format; they do not constitute independent human annotation or establish every scientific claim. The authors are responsible for the manuscript and the AI-assisted artifacts.

ETHICS STATEMENT

The subjects are audience members, not only public-facing creators. Processing their images and contextual records can create risks of unwanted profiling, reidentification, and stereotyping. The recorded workflow sends profile evidence to a third-party teacher through its native interface. The experiments do not establish participant consent, a legal basis for reuse, or an institutional-review determination; no such approval is claimed.

A fixed exclusion mask prevents the release of sensitive-field predictions, and the teacher instructions prohibit inferring excluded properties through proxies. This constrains outputs rather than certifying the model's internal reasoning. We report aggregates and withhold handles, identifying examples, proprietary field names, and individual predictions. Internal pseudonyms are not anonymization. Any deployment requires appropriate data-access rights, purpose limitations, and institutional review where applicable. The proposed measurement application does not justify high-stakes decisions about individuals.

REPRODUCIBILITY STATEMENT

The method and accounting definitions appear in Sections 3–4. Experimental counts and their denominators are stated with the pilot results. An anonymous NumPy supplement implements full-item aggregation, residual scoring, fixed-prior construction, and the censored objective, with explicitly synthetic arithmetic tests. A private correspondence check matches all 9,524 saved logits in each of four frozen comparison arms, with maximum absolute discrepancy 8.88×10^{-16} and exact support masks. This checks numerical correspondence, not predictive quality. Exact source snapshots, model and prediction fingerprints, output validation, and usage receipts are retained privately. Restricted profiles and the application-specific field vocabulary are not released. Consequently the supplement does not reproduce the restricted-data experiments or supply a complete training and deployment pipeline. The distinction between measured results, diagnostics, and unexecuted evaluation requirements is retained throughout.

REFERENCES

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023. URL https://arxiv.org/abs/2305.05176.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023. doi: 10.1073/pnas.2305016120. URL https://arxiv.org/abs/2303.15056.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2127–2136, 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Lunyiu Nie, Zhimin Ding, Erdong Hu, Christopher Jermaine, and Swarat Chaudhuri. Online cascade learning for efficient inference over streams. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 38071–38090. PMLR, 2024. URL https://proceedings.mlr.press/v235/nie24a.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL https://arxiv.org/abs/2103.00020.

Guillem Ramírez, Matthias Lindemann, Alexandra Birch, and Ivan Titov. Cache & distil: Optimising API calls to large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11838–11853. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.704. URL https://aclanthology.org/2024.findings-acl.704/.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general chinese embeddings. *arXiv preprint arXiv:2309.07597v5*, 2024. URL https://arxiv.org/abs/2309.07597v5.

A Local model specification

The image checkpoint is `timm/vit_base_patch32_clip_224.openai`; each full frame is resized by its longest dimension, padded to 512×512, and represented by a normalized 512-dimensional vector. The text checkpoint is `BAAI/bge-small-en-v1.5`; the exact stored evidence is tokenized without truncation or a retrieval-instruction prefix, divided into contiguous chunks of 510 content tokens plus special tokens, and represented by normalized 384-dimensional CLS vectors. Field–value candidate embeddings average two schema-only text templates, encoded with CLIP's text tower for the image space and BGE for the text space. Candidate templates contain field and value names, not the complete field definitions. This compression can lose information despite retaining every input item.

For modality m, item vector x_{it}^m, candidate vector v_{jk}^m, and normalized mean candidate vector q_j^m, the aggregation weights at temperature $\tau = 0.1$ are

$$\alpha_{ijt}^m = \frac{1}{2N_i^m} + \frac{1}{2}\,\mathrm{softmax}_t\left((x_{it}^m)^\top q_j^m/\tau\right), \qquad c_{ij}^m = \sum_t \alpha_{ijt}^m x_{it}^m. \tag{4}$$

Every retained item has positive weight. The low-rank residual is

$$r_{ijk} = \frac{(c_{ij}^I A_I)^\top (v_{jk}^I B_I) + (c_{ij}^T A_T)^\top (v_{jk}^T B_T)}{2\sqrt{8}} + \beta^\top \phi_{ijk}. \tag{5}$$

The image matrices have shape 512×8, the text matrices 384×8, and β has eight entries. The eight statistics comprise, for each modality, normalized-context–candidate cosine, mean item cosine, candidate-centered temperature-scaled log-mean-exp item cosine, and context-cosine margin against the best competing candidate. The bilinear residual uses the unnormalized weighted context. For the log-mean-exp statistic, let $a_{ijk} = \tau \log[N_i^{-1} \sum_t \exp(x_{it}^\top v_{jk}/\tau)]$; its centered value is $a_{ijk} - |\mathcal{V}_j|^{-1} \sum_{\ell \in \mathcal{V}_j} a_{ij\ell}$. The singleton-vocabulary margin is zero. This yields 14,344 trainable parameters shared across fields.

Reliability weights are 1 for admitted study-consensus and current known labels, and 0.25 for known labels from a single archival database row. The study-consensus inclusion criterion is at least four of five responses; it is not a weight of 0.8. Explicit unknown and unavailable cells contribute no supervised loss. The corrected logits are $z_{ijk} = \log \pi_{jk} + r_{ijk}$, where π_{jk} is the training-only reliability-weighted empirical class prior. Zero-support classes are excluded without smoothing, and a field lacking supported classes returns UNKNOWN. Training uses Adam (Kingma & Ba, 2015), learning rate 0.03, moments $(0.9, 0.999)$, epsilon 10^{-8}, Gaussian matrix initialization with standard deviation 0.02, zero statistics weights, seed 20260926, and $\lambda_2 = 10^{-4}$. Full-source gradients are accumulated in deterministic eight-profile batches. The 40-update comparison snapshot is fixed in advance. Continued source-only training checks the best objective every ten updates after update 40, stopping after two consecutive checks with relative best-objective improvement at most 10^{-4} over the preceding 20 updates, or at 200 updates. The observed run stops at update 60; this later checkpoint does not replace the predeclared source-40 comparison. Warm and cold current adaptation each use exactly 20 updates with a fresh optimizer and current-only priors and support.

A.1 Compute and model footprint

The local experiments use an Apple M4 Pro with 24 GiB unified memory and 14 logical CPU cores. Numeric workers use one CPU thread; the encoder also uses one PyTorch inter-operation thread. Feature extraction is paced to 20% of one core and corrected source training to 35%, with thermal backoff. These are measured process-duty targets, not instantaneous host-wide limits. No GPU is used. Software versions are NumPy 2.0.2, PyTorch 2.8.0, TorchVision 0.23.0, OpenCLIP 3.3.0, and Transformers 4.57.6.

The CLIP visual checkpoint has 87,849,216 stored floating elements, while BGE has 33,360,000 plus a 512-element integer position buffer. The full downloaded CLIP checkpoint has 151,277,313 floating elements overall, including the text tower used to construct the fixed candidate cache but unused during subsequent profile-image feature extraction. Frozen checkpoints, tokenizers, semantic caches, profile features, and optimizer history are excluded from the 14,344-parameter head count. A prior-aware head stores 114,752 bytes of float64 fitted weights and 42,858 bytes of fixed priors and masks; its NumPy parameter archive is 159,416 bytes before JSON metadata. This is not the footprint of the complete inference system.

Corrected source training completes 60 updates in 1,512.197 wall seconds and 525.545 process-tree CPU seconds, with 612.073 MB peak sampled resident memory. An earlier memory-limited attempt fails before its first update and remains in the cost record. Independent replay of the selected model reproduces its known-label cross-entropy (0.4572877564) and regularized objective (0.4601391021). Numerical warnings from the original NumPy/Accelerate computation are retained; independent replay verifies the saved final values and current prediction outputs, not every intermediate training gradient. These training checks do not measure generalization.

B EVALUATION AND RELEASE CRITERIA

The complete evaluation requires independent annotations; field-level precision and recall; micro and macro aggregation; coverage and calibration; per-profile and per-community uncertainty; total training, calibration, preprocessing, and inference expense; measured throughput including failed cases; and matched-budget comparisons with fixed distillation, fixed cascades, and random acquisition. A policy is released only after its acceptance rule is frozen and its quality threshold is met on data not used to fit or choose that policy. Returning UNKNOWN or escalating to the teacher remains preferable to asserting an unsupported classification.

C COST INTERPRETATION

Equation 2 assumes equal teacher cost across profiles only for exposition. In a heterogeneous workload, replace $M(q_0 - q_1)c_T$ by the sum of avoided teacher expenses on the actual routed profiles. Input resolution, output length, caching, retries, and batching can all change those expenses. Counted-token reduction is reported separately from dollar-cost reduction, and both are reported separately from predictive quality. A million-profile arithmetic extrapolation is not a million-profile experiment.