Sciev: Typed Scientific Decisions Without Generation

Angel Luis Robles-Fernández[1]

[1]Kansas Biological Survey, University of Kansas, Lawrence, KS, USA,
a.l.robles.fernandez@gmail.com

September 27, 2026

Abstract

Scientific workflows repeatedly require bounded judgments: selecting an answer, assessing whether evidence supports a claim, or rating a proposed response. We present Sciev, an open System-One-style decision layer that produces these typed outputs without text generation. Specialist heads read the hidden states of a frozen LLaDA-8B masked-diffusion backbone, using one forward pass per decision and deterministic option ordering. We study an unadapted backbone and one archival, experimentally adapted comparator under the same head-training recipe and three head seeds. On a complete-passage scientific battery with constructed reference labels, the frozen arm reaches mean accuracies of 97.0%, 92.5%, and 85.9% for selection, binary verification, and ordinal scoring. However, its selection head still agrees with the reference on 87.5% of items when the passage is removed, exposing substantial evidence-independent predictability. The adapted comparator improves external SciFact binary verification from 48.3% to 70.9% while reducing internal ordinal accuracy to 77.9%; GPQA remains weak for both arms. Calibration on the internal distribution does not ensure calibration after transfer. We provide a fingerprinted, machine-readable account of 60 completed evaluation reports, including controls and per-seed values. The results support typed, non-generative interfaces and evidence-controlled evaluation, but do not establish general scientific reasoning, semantic grounding from accuracy alone, or deployment error guarantees.

1 Introduction

A scientific workflow contains many decisions that do not require an open-ended answer: which candidate matches an observation, whether a passage supports a claim, which tool should receive a request, or which rubric level best describes a response. Language-model pipelines often obtain these judgments by generating a label or a structured response and then interpreting it. Constrained decoding can improve output validity, but the underlying task can also be formulated directly as a conditional distribution over an explicit set of alternatives. The distinction concerns the interface and inference procedure, not a guarantee that the resulting judgment is correct.

System-One interfaces package such judgments as typed questions over a supplied state. Jev provides a proprietary implementation [2]; open systems include encoder-based decision layers, autoregressive readouts, and diffusion answer-slot readers [3, 4, 5, 6]. This interface is useful for scientific workflows because the alternatives, returned probabilities, and subsequent routing policy can be recorded separately from any generated explanation. Classification and ordinal prediction are established problems; our aim is to examine a particular open implementation and the evidence needed to interpret its scientific performance.

SCIEV uses LLaDA-8B-Instruct [1] as a frozen feature extractor, with specialist heads for selection, binary support judgments, and ordinal ratings. The backbone was pretrained as a masked-diffusion language model, but our released readout consumes a fully observed sequence: it neither inserts answer masks nor runs a denoising generation loop. We ask whether this layer can make useful scientific decisions and, crucially, whether benchmark accuracy indicates dependence on the supplied evidence.

The study has three contributions:

1. An open, non-generative decision implementation with a shared input contract, specialist span readouts, and canonical option presentation. For options with distinct token encodings, presentation permutations reduce to an identical model input.

2. An evidence-controlled scientific evaluation that separates accuracy against constructed reference labels from agreement after evidence removal or replacement. High selection accuracy persists without the passage, while scoring exhibits greater evidence sensitivity.

3. A complete, cohort-explicit comparison of unadapted and experimentally adapted features across three head-training seeds. The observed transfer and calibration effects are task-dependent. Machine-readable summaries, artifact fingerprints, and analysis code accompany the manuscript.

These contributions are deliberately narrower than a claim of a generally reliable scientific reasoner. Our internal labels are partly heuristic, the external evaluations informed development, and only one adapted backbone was produced. We report these limitations as part of the result rather than treating interface compatibility or high internal accuracy as scientific validation.

2 Related Work

Typed decision interfaces. Jev [2] motivates the System-One wire interface. Laya and cbjev [3, 4] use encoder-based decision models; Kev [5] implements related readouts on Qwen-family backbones. OpenJev [6] already uses DiffusionGemma to read masked answer slots without free-form generation. Accordingly, SCIEV is not the first diffusion-based System-One implementation. Its focus here is trained, type-specific readouts over observed LLaDA features and their evaluation on scientific reference tasks and evidence controls. We make no protocol-matched performance ranking against these systems.

Diffusion backbones and frozen readouts. Discrete and masked diffusion provide non-autoregressive language-modeling objectives [7, 8, 9, 10, 11]; LLaDA and Dream extend this family to large pretrained language models [1, 12]. Reading frozen intermediate representations relates to probing and representation reuse [13]. Our inference procedure uses the resulting bidirectional representations, not diffusion sampling. The adapted comparator draws on the broader ideas of domain-adaptive pretraining and low-rank updates [14, 15], but its historical loss differs from the native LLaDA estimator; we state the implemented procedure explicitly in Section 4.4.

Calibration and selective prediction. Temperature scaling adjusts predictive distributions using held-out labels [16, 17]. Selective prediction studies the trade-off between coverage and error [18, 19], while conformal methods obtain particular guarantees under additional assumptions [20]. We distinguish probability estimates, empirical calibration, and retrospective oracle coverage. None is an automatic guarantee of error control on a new scientific domain, and we do not evaluate a prospective deployment policy here.

Evidence dependence and scientific verification. Hypothesis-only baselines and annotation-artifact analyses show that a task may be solvable without its intended evidence [21, 22]. Our controls apply that principle to a typed decision layer. SciFact situates verification in scientific evidence [23, 24], whereas GPQA probes difficult scientific question answering without a supplied passage [25]. Our synthetic ordinal task is narrower than human-aligned response evaluation with generative judges [26, 27]. Finally, canonical presentation addresses option-order sensitivity [28]; it does not address paraphrase sensitivity, invalid alternatives, or incorrect evidence.

3 Decision Layer

3.1 Typed questions and outputs

A request supplies a state s, instructions q, and a finite collection of alternatives O. The state and instructions may be text or deterministically serialized JSON. The interface supports:

- **choice**: a distribution over named alternatives and a selected option. The scientific training task uses four candidate answers.
- **noul**: the interface's binary decision type. In this study it represents support versus lack of support for a proposed answer or claim, and returns the estimated probability of support.
- **score**: a distribution over ordered rubric levels. The API returns the expected level $\sum_k kp_k$ as well as the distribution; our internal task uses levels $0, 1, 2$. Accuracy is evaluated on the argmax level, not on a rounded expectation.

The API's concentration-style `confidence` field is distinct from maximum option probability and is not $P(\text{correct})$. Calibration and selective metrics below use $c = \max_k p_k$.

3.2 Observed-sequence readout

Context, instructions, and alternatives are encoded as one observed token sequence. Frozen LLaDA-8B-Instruct produces contextualized token representations. For binary verification and scoring, a mean-pooled final-layer representation of each option span is passed through a shared MLP scorer. For choice, a learned query pools each option span separately at four hidden-state outputs; their concatenation is passed to a shared scorer. The layer indices are $(-1, -9, -17, -25)$, corresponding to outputs $(32, 24, 16, 8)$ when the embedding output is indexed zero. The choice head has 67,166,209 parameters; each MLP has 16,793,601. Heads are separate across decision types, but shared across the options of a decision.

Let $z_k = f_\theta(h_k)$ be the score of option k. Inference returns

$$p_k = \frac{\exp(z_k/T)}{\sum_{j=1}^{K} \exp(z_j/T)}, \qquad T > 0. \tag{1}$$

All alternatives are scored from the same backbone forward pass. A request containing several questions is processed as several decisions; it is not a single shared forward for the whole request. The output is serialized from validated alternatives and finite probabilities, with no generated answer string to parse. Incorrect decisions remain possible.

3.3 Canonical option presentation

The input builder sorts option token sequences lexicographically, records their original identities, and remaps the output distribution back to those identities. If all option encodings are distinct, any permutation of their presentation produces the same sorted sequence. Thus the encoded input is invariant and the identity-aligned output is equivariant under presentation permutation, assuming a fixed deterministic computation. Indistinguishable option encodings are rejected rather than assigned an arbitrary gold identity.

This is an input-construction property, not an empirical demonstration of semantic robustness. Renaming options, changing their descriptions, or paraphrasing the question can change the encoded sequence. Our zero-flip diagnostic under canonical ordering should therefore not be interpreted as resistance to such changes.

3.4 Head optimization and temperature fitting

The main loss is cross-entropy against the reference option. For the score head, we add a cumulative-probability ordinal term related to rank-consistent ordinal learning [29]:

$$\mathcal{L}_{\text{score}} = -\log p_g + \sum_{j=1}^{K-1} \text{BCE}\left(\sum_{k=j}^{K-1} p_k, \mathbf{1}[g \geq j]\right), \tag{2}$$

with $T = 1$ during training. This is an auxiliary loss on a softmax distribution, not an implementation of independent CORAL output heads.

Both backbone arms use the same `scientific-v1` head recipe: AdamW, learning rate 3×10^{-4}, weight decay 0.01, linear warmup over 200 example steps, accumulation one, canonical ordering, and one presentation per training example. Choice receives 2,000 example updates; noul and score receive 3,000 each. Examples are traversed in shuffled order, restarting when necessary. An example update is not an epoch. The backbone is frozen throughout head training in both arms.

For each checkpoint, a positive scalar temperature minimizes NLL on the corresponding scientific dev split using a coarse grid followed by log-scale one-dimensional refinement. External benchmark labels are not used for this fit. All reported predictive metrics use these dev-fitted temperatures. Temperature changes probabilities, but not the argmax decision.

4 Study Design and Data

4.1 Scope of the evidence

We freeze a complete archival study rather than mixing successive candidate families. The main comparison contains 42 reports: two backbone arms, three head-training seeds (7331–7333), and seven evaluation tasks. Twelve additional reports contain empty/shuffled evidence controls for seed 7331, and six contain a stricter-pool choice variant. These 60 reports reference 18 head checkpoints and 22 distinct encoded datasets. Their full-file SHA-256 fingerprints and compact metrics are provided in the ancillary evidence ledger.

The frozen arm is denoted `fr`; the adapted comparator is `da`. The public v0.2.2 release distributes the seed-7331 frozen heads as `fr_choice.pt`, `fr_noul.pt`, and `fr_score.pt` [30]. The earlier `c_*` family is not the released family evaluated in our main tables. Incomplete later prediction runs, QA-specialized pilots, and legacy defective-input ablations are not pooled into this study. Appendix C explains the exclusion criteria.

4.2 Complete-passage scientific battery

The battery starts from 3,339 supplied scientific question–answer records with passage identifiers. A lexical consistency filter retains 3,150 records: content-word recall of an answer against its passage must be at least 0.5, and its numeric literals must occur in the passage. Questions are limited to 400 characters and answers to 8–512 characters, with nonempty text fields. These checks constrain surface consistency; they do not establish factual correctness. The input collection includes factual, numerical, causal, multihop, negation, and definitional question tags, but we do not use those tags as validated measures of reasoning ability.

Connected source identifiers and normalized passage content define split groups. Before encoding exclusions, the split contains 500/115/153 groups and 2,029/493/628 QA records for train/dev/evaluation. Each record can yield several decisions. Choice pairs the reference answer with three alternatives drawn from same-passage answers, numerical perturbations, and, when necessary, other passages in the same split. Noul treats the source answer as supported and heuristic alternatives as unsupported. Score assigns 2 to the source answer, 1 to a related-but-altered answer, and 0 to a cross-passage answer.

These are *constructed reference labels*. A same-passage swap need not be false for every question, a foreign answer need not be irrelevant, and a source answer need not be scientifically correct. The score labels are particularly narrow proxies for the generation strategy, not expert ratings of arbitrary scientific responses. Negative pools are split-local to avoid deliberately sharing alternatives across evaluation boundaries.

The `systemone-v2` encoding accepts the complete supplied context or excludes the decision. It allows up to 960 context/instruction tokens, 120 per option, and 2,048 in the model sequence. The manifest records 8,576 *decision-level* overflow exclusions, not 8,576 distinct passages, as well as 189 lexical-filter exclusions and 13 unavailable related negatives. The resulting counts appear in Table 1. The selected encoded files contain no truncated contexts/options, colliding option encodings, or conflicting effective inputs. Isolation is at the declared source-group and normalized-passage level; document-level disjointness and paraphrase leakage are not established.

Decision type	Train	Dev	Evaluation
choice	1307	300	448
noul	3649	814	1200
score	3685	827	1218

Table 1: Decision counts after filtering and complete-input encoding. Dev is used for temperature fitting. Counts differ across types because candidate availability and complete-context budgets differ.

4.3 External evaluations

GPQA [25] supplies a scientific question and four candidate answers, without an evidence passage. It therefore probes parametric question answering, not evidence-conditioned verification. The archived conversion retains 441 main and 193 diamond questions. Every retained row has four distinct options and complete encoded input. Diamond is a subset of main and is not treated as an independent replication. A later 440-item main conversion removes one 923-token context under a tighter budget; the other 440 effective inputs are identical. We retain the explicitly identified 441-item cohort for all six matched heads rather than mixing protocols.

SciFact [23] supplies claim–abstract pairs. Its archived conversion retains 332 of 340 input pairs for two separate tasks: binary SUPPORT versus CONTRADICT/NEI using the noul head, and nominal three-way verdict selection using the choice head. Both tasks condition on supplied abstracts; evidence retrieval and sentence-level rationale extraction are outside this evaluation. The class counts are 134 SUPPORT, 69 CONTRADICT, and 129 NEI. Consequently the binary majority reference is 198/332 (59.6%), whereas the nominal majority reference is 134/332. The three-way task is *not* ordinal scoring. Later 304-item clean conversions are different cohorts and are not substituted into these tables.

These benchmark files are not direct head-training or temperature-fitting inputs. However, external performance informed candidate development and release decisions. All external results are therefore selection-informed, exploratory estimates, not an untouched confirmatory test. We cannot exclude benchmark-related text from backbone pretraining or comprehensively establish its absence from the adaptation corpus.

4.4 An archival adapted comparator

The adapted arm uses one existing LoRA adapter on LLaDA-8B-Instruct, trained on a mixture of science-derived text collections. The recorded configuration uses rank 16, scaling parameter 32, dropout 0.05, AdamW learning rate 2×10^{-4}, batch size 16, and a 1,024-token cap. The final nominal step is 5,000. The run warm-starts at step 2,000 with a new optimizer and without restoring random-number-generator state; it is not an exact training-state resume. The nominal budget is at most 81.92 million input token slots, including padding, not a measured count of unique corpus tokens.

The historical implementation samples a mask rate $t_b \sim U(0, 1)$ for each sequence and uses the recorded LLaDA mask token 126336. Let M contain the masked token positions, ℓ_{bi} their token cross-entropies, and $\bar{t} = B^{-1} \sum_b t_b$. The implemented loss is

$$\mathcal{L}_{\text{archived}} = \frac{|M|^{-1} \sum_{(b,i)\in M} \ell_{bi}}{\max(\bar{t}, 10^{-3})}. \tag{3}$$

Padding positions are not separately excluded from this masked-token objective. Equation 3 is **not the native** per-token importance-weighted LLaDA estimator, which divides each masked loss by its own mask probability before normalization over all token slots. We preserve this distinction rather than retrospectively describing the adapter as a standard native-objective DAPT replication.

The adapter is merged into the backbone before head training; both arms then freeze their respective backbone weights. Thus the comparison holds the head architecture, data, optimization recipe, and head seed fixed while changing one particular set of adapted features. It does not estimate the effect of canonical DAPT in general, compare optimized adaptation recipes, or measure variation across independently trained adapters.

4.5 Evidence controls and stricter alternatives

For seed 7331, an empty control removes only the recorded evidence span while preserving the question, candidate answers, and proposed response. A shuffled control substitutes a donor passage from a different source group in the same evaluation split. These controls are derived from the complete-passage `systemone-v2` text, not from the earlier truncated battery.

Controls retain the original reference label but do not assign new truth labels to changed evidence. We therefore call their metric *reference agreement*, not verification accuracy. Empty controls retain all original decisions; shuffling excludes donor-overflow cases, leaving 447/1,146/1,172 decisions

for choice/noul/score. Neither low nor high agreement alone identifies whether a model correctly interprets the altered evidence.

A stricter choice variant excludes cross-passage distractor fill and uses only same-passage swaps and numerical perturbations. It retains 303 evaluation questions; 733 QA records across the builder's splits lack enough eligible alternatives. Heads are not retrained for this variant. Because its eligible question subset differs from the original one, a difference between the two batteries is not an isolated causal estimate of distractor difficulty.

4.6 Metrics and uncertainty

We report accuracy, natural-log negative log-likelihood $-n^{-1}\sum_i \log p_{i,g_i}$, and multiclass Brier score $n^{-1}\sum_i \sum_k (p_{ik} - \mathbf{1}[k = g_i])^2$, without dividing by the number of classes. A zero gold probability is recorded explicitly rather than silently assigned a finite NLL. ECE uses ten equal-width confidence bins:

$$\mathrm{ECE}_{10} = \sum_{j=1}^{10} \frac{|B_j|}{n} \left| \bar{c}_{B_j} - \overline{\mathbf{1}[\hat{y} = g]}_{B_j} \right|. \tag{4}$$

Bins are right-closed, with zero included in the first bin. Reliability-bin summaries and additional fixed-label/ordinal metrics are retained in the ancillary records. ECE is estimator- and sample-dependent; NLL and Brier provide complementary proper scoring-rule summaries.

For the main comparison we report the mean and *sample standard deviation* over three head seeds. These are not confidence intervals for population accuracy and do not include adaptation-seed uncertainty. The archival reports do not provide complete per-item predictions for all arms, so we do not claim paired item-level significance or bootstrap intervals. Each evaluation report uses two presentations per decision for an order diagnostic; the predictive metric uses the first presentation. Canonical ordering yields zero flips in these reports. Recorded loop times and forward counts are supplied, but are not a controlled hardware-normalized latency comparison.

We also retain a retrospective oracle coverage metric: the largest fraction accepted by a maximum-probability threshold with at most 5% observed error on the same labeled evaluation set. Equal-confidence groups cannot be split. This is not a dev-selected threshold or a deployment guarantee; it appears only as a diagnostic in Appendix B.

5 Results

5.1 High reference accuracy, uneven transfer

Table 2 shows high agreement with internal reference labels: the frozen heads reach 97.0%, 92.5%, and 85.9% mean accuracy. This establishes that the typed readout can learn the constructed tasks; it does not establish the validity of their references or dependence on the intended evidence.

Transfer differs sharply by task. On SciFact binary verification, the frozen arm's 48.3% lies below the 59.6% majority reference, whereas the adapted comparator reaches 70.9%. The mean gain is 22.6 percentage points and has the same sign for all three head seeds (Appendix A). This is the largest consistent transfer difference in the selected study. For nominal three-way SciFact selection, the mean adapted accuracy instead falls below the frozen result, with substantial head-seed variation.

The internal ordinal task also favors the unadapted features on average: 85.9% versus 77.9%. Its large seed spread prevents an interpretation based on a single especially favorable or unfavorable head. These results characterize task-specific interactions between a fixed adapter and trained readouts, not a universal benefit or cost of adaptation.

Evaluation	n	Uniform	Majority	Frozen	Adapted
Scientific choice	448	0.250	–	0.970 ± 0.017	0.961 ± 0.011
Scientific verification	1200	0.500	0.667	0.925 ± 0.006	0.913 ± 0.007
Scientific score	1218	0.333	0.336	0.859 ± 0.075	0.779 ± 0.109
GPQA main	441	0.250	–	0.295 ± 0.014	0.288 ± 0.018
GPQA diamond	193	0.250	–	0.307 ± 0.012	0.297 ± 0.011
SciFact binary	332	0.500	0.596	0.483 ± 0.023	0.709 ± 0.020
SciFact three-way	332	0.333	0.404	0.479 ± 0.018	0.444 ± 0.062

Table 2: Matched-study accuracy, mean ± sample SD over head seeds 7331–7333. Uniform is $1/K$; majority is the largest observed class fraction for tasks with fixed semantic labels, not a classifier selected on dev. The adapted arm uses the historical procedure in Section 4.4. GPQA main and diamond overlap.

GPQA main accuracy is 29.5% for the frozen arm and 28.8% for the adapted arm, against 25% uniform guessing. This is weak performance for difficult scientific questions. The heads were trained on passage-conditioned decisions, not on a parametric QA curriculum. The observed scores do not establish a knowledge ceiling for LLaDA or show that the heads extract all information available through other inference protocols.

5.2 Calibration does not transfer uniformly

Evaluation	Arm	ECE	NLL	Brier
Scientific choice	fr	0.018 ± 0.006	0.091 ± 0.042	0.047 ± 0.024
	da	0.014 ± 0.003	0.104 ± 0.026	0.059 ± 0.016
Scientific verification	fr	0.013 ± 0.005	0.175 ± 0.009	0.108 ± 0.006
	da	0.018 ± 0.002	0.196 ± 0.009	0.123 ± 0.007
Scientific score	fr	0.037 ± 0.028	0.348 ± 0.159	0.205 ± 0.097
	da	0.056 ± 0.035	0.451 ± 0.196	0.277 ± 0.118
GPQA main	fr	0.236 ± 0.016	1.637 ± 0.058	0.844 ± 0.018
	da	0.278 ± 0.072	1.792 ± 0.177	0.888 ± 0.051
GPQA diamond	fr	0.223 ± 0.017	1.594 ± 0.088	0.830 ± 0.039
	da	0.271 ± 0.087	1.741 ± 0.197	0.872 ± 0.062
SciFact binary	fr	0.280 ± 0.037	0.744 ± 0.021	0.551 ± 0.019
	da	0.134 ± 0.025	0.615 ± 0.001	0.423 ± 0.001
SciFact three-way	fr	0.262 ± 0.032	1.563 ± 0.314	0.745 ± 0.053
	da	0.375 ± 0.199	2.153 ± 1.027	0.895 ± 0.228

Table 3: Dev-temperature-scaled predictive metrics, mean ± sample SD across the same three head seeds. Lower is better for all columns. Temperatures are fitted on scientific dev data, not on these external benchmark labels. Low ECE on reference-labeled internal tasks is not evidence of universal calibration or scientific correctness.

The frozen arm's mean internal ECE is 0.018, 0.013, and 0.037 for choice, noul, and score. In contrast, its GPQA-main ECE is 0.236 and its SciFact-binary ECE is 0.280 (Table 3). The adapter improves both accuracy and the reported calibration summaries on binary SciFact, but not across all external tasks. For GPQA main, uniform four-option probabilities have NLL $\log 4 \simeq 1.386$ and Brier score 0.75; both arms have worse mean proper scores despite accuracy above uniform guessing. This distribution dependence matters for routing: a useful dev temperature in one reference task is

not sufficient justification for accepting high-confidence decisions in a different domain.

5.3 Evidence-independent selection survives stronger distractors

Figure 1 and Table 4 compare intact inputs with passage removal and replacement, retaining the original references.

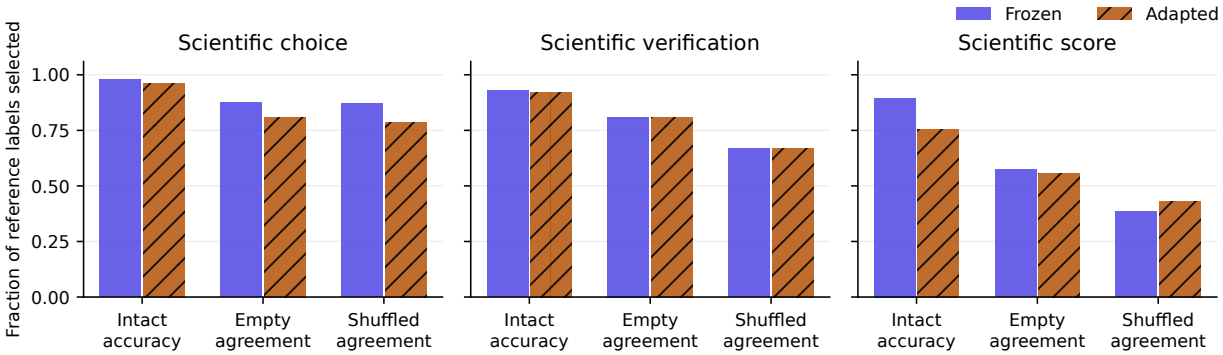

Figure 1: Seed-7331 evidence controls. Intact inputs have reference accuracy; changed-evidence conditions have agreement with the *original* reference, not relabeled truth. Shuffled cohorts exclude donor-overflow cases, so changes are not paired effect estimates on identical complete cohorts. Exact counts and values are given in Table 4.

Task	Arm	Intact acc.	Empty agree.	Shuffled agree.
Scientific choice	fr	0.9799	0.8750	0.8725
Scientific choice	da	0.9598	0.8058	0.7852
Scientific verification	fr	0.9308	0.8100	0.6693
Scientific verification	da	0.9200	0.8067	0.6667
Scientific score	fr	0.8957	0.5747	0.3848
Scientific score	da	0.7521	0.5550	0.4300

Table 4: Evidence controls for the complete-passage battery, seed 7331. Intact/empty/shuffled counts are 448/448/447 for choice, 1,200/1,200/1,146 for noul, and 1,218/1,218/1,172 for score. Internal noul's intact majority reference is 2/3, not the different SciFact majority. Agreement after evidence changes is a diagnostic rather than a correctness measure.

The strongest qualification to the internal choice result is its 87.5% reference agreement without the passage; the adapted arm retains 80.6%. Much of this task can therefore be solved without the supplied evidence. Teacher-answer style, question–answer coherence, topical differences among alternatives, and pretrained knowledge are plausible contributors. The construction makes such shortcuts possible, but these controls do not separately identify their contributions.

Noul and score are more sensitive to the intervention. Under shuffled evidence, score agreement is 38.5% and 43.0% for frozen and adapted heads. This supports evidence sensitivity, not the stronger claim that their ordinal judgments correctly use evidence. The proxies distinguish synthetic answer-generation strategies, and altered passages have not been independently relabeled.

Even when distractors are limited to same-passage swaps and numeric perturbations, empty-evidence agreement remains 79.5% and 76.6% (Table 5). Answers to *other* questions in the same passage may remain distinguishable by question–answer coherence. We therefore do not treat this

Arm	n_{intact}	n_{empty}	$n_{shuffle}$	Intact acc.	Empty agree.	Shuffled agree.
fr	303	303	302	0.9142	0.7954	0.8245
da	303	303	302	0.8680	0.7657	0.7219

Table 5: Stricter-pool choice evaluation with the existing seed-7331 heads. Counts are explicit because eligible subsets differ from the standard battery. The altered-evidence columns remain reference agreement.

construction as a validated hard scientific QA benchmark. Same-question alternatives whose validity changes with controlled evidence, accompanied by independently checked labels, are needed to test that stronger claim.

6 Discussion

Interface validity and scientific validity are different. A typed layer exposes a bounded decision and its distribution without a free-form decoding step. Canonicalization removes one source of presentation variation. Neither property ensures that the alternatives are scientifically meaningful, that their labels are correct, or that the chosen option follows from the provided evidence. The high internal accuracies and their evidence controls illustrate why these axes must be evaluated separately.

A useful result can be a constraint on interpretation. The persistence of selection agreement under empty and mismatched passages is not evidence that scientific verification has been solved. It identifies what the present battery cannot establish. Similarly, the SciFact binary improvement is a concrete observation about one adapted representation and three readouts, while the GPQA and ordinal results prevent a uniform adaptation-success narrative. Reporting all tasks and seeds is more informative than selecting a favorable endpoint.

Calibration should support, not replace, task validation. Post-hoc temperatures and retrospective confidence curves can guide subsequent validation. They do not make a new scientific task safe to automate. A prospective system would need a fixed policy selected without the final evaluation labels, verified model/input compatibility, representative expert-labeled validation, and monitoring under distribution shift. Our serving implementation currently reports calibration as unverified; this paper does not certify its confidence field or acceptance behavior for deployment.

Next scientific tests. The most useful extensions are not simply larger training pools. They include expert-checked, same-question distractors; evidence counterfactuals with new truth labels; document-held-out scientific corpora; independent adapter runs; properly implemented alternative adaptation objectives; and matched external baselines on identical inputs. Better parametric QA extraction and training remain separate questions from passage-grounded decision quality.

7 Limitations

The internal data are small, curated, and reference-labeled through a partially heuristic construction. Surface consistency filters are not semantic validators. Excluding long inputs changes the sampled task distribution, and the reported results do not cover the excluded cases. Passage/source-group

isolation is not complete document isolation, and paraphrases or benchmark-adjacent pretraining text may remain. Provenance of the original QA generator and all upstream corpus licensing is not fully established by the available manifests.

Only three head seeds and one archival adapted backbone are evaluated. The adapter used a non-native loss normalization, included masked padding, and warm-started without optimizer/RNG restoration. Its results are not evidence about an optimal or standard DAPT procedure. We do not infer catastrophic forgetting from excluded full-fine-tuning artifacts. No complete clean factorial ablation isolates the effects of pooling, specialization, canonicalization, and ordinal loss in this study.

Evidence controls use one head seed and retain old references rather than annotating changed-evidence truth. Donor-overflow exclusions and the stricter pool's different eligibility prevent causal comparisons of their aggregate changes. We do not distinguish memorized knowledge from option-side cues, and we do not present complete paired item-level intervals or prospective selective risk measurements. Multiple questions can share a passage; treating all such decisions as independent observations would be unjustified.

External results are selection-informed. The published cohorts are converted subsets, not official leaderboard protocols; main and diamond GPQA overlap. The scientific heads saw four choice options and a three-level score rubric, so flexible API arity does not demonstrate transfer to arbitrary option counts or rubrics. The study contains no matched ranking against proprietary models or other open System-One systems and no controlled end-to-end latency benchmark. We do not establish that an 8B diffusion backbone is necessary or more efficient than smaller encoders for these tasks.

8 Conclusion

Sciev demonstrates a concrete open route from scientific state and typed questions to structured decisions without generation. Its specialist heads can achieve high accuracy on constructed scientific references, but those numbers overstate what the same experiments establish about dependence on provided evidence. The archival adaptation contrast also shows why transfer and calibration must be evaluated per task rather than inferred from a single internal score. The practical contribution is a typed implementation and an auditable evaluation that keeps structural guarantees, reference accuracy, evidence sensitivity, and prospective reliability distinct.

Code, Data, and Artifact Availability

Code and frozen decision-head checkpoints are available in the public Sciev repository and v0.2.2 release [30]; the backbone is obtained separately from the LLaDA authors [1]. Code and released heads use Apache-2.0, while the backbone and source datasets retain their own terms. The ancillary files accompanying this manuscript provide all 60 selected report summaries, per-seed metrics, dataset and checkpoint fingerprints, training configurations, and code to regenerate the statistical tables and control figure. They contain no benchmark question text. After placing the ancillary files in one directory, the statistical ledger can be regenerated using only Python's standard library:

```
python paper_analysis.py --evidence evidence.json \
    --out reproduced_results.json
```

Matplotlib is needed only to redraw the figure; the observed analysis environment and additional commands are recorded in `analysis_environment.json`.

This supports independent reproduction of the *reported arithmetic*, not a claim of fully self-contained retraining. The original scientific QA collection is not redistributed here, and the public

release does not bundle the experimental adapter needed to rerun the adapted arm. The ancillary ledger records its checksum and configuration but not its weights. Full historical GPU/software environments and exact backbone-revision pins are also not available for every run. The manuscript's evidence ledger is the authoritative mapping for its tables; older release documentation and candidate summaries must not be substituted for it.

Acknowledgments and Tool Use

Computations used the University of Kansas HPC cluster. Generative AI tools, including Devin, assisted software development, manuscript preparation, and artifact auditing. Numerical claims in this version are linked to recorded experimental outputs and deterministic summary calculations. The human author is responsible for the scientific claims, references, interpretation, and submission; these tools are not authors.

A Per-Seed Accuracy

Evaluation	Arm	7331	7332	7333
Scientific choice	fr	0.9799	0.9509	0.9799
Scientific choice	da	0.9598	0.9732	0.9509
Scientific verification	fr	0.9308	0.9258	0.9192
Scientific verification	da	0.9200	0.9117	0.9067
Scientific score	fr	0.8957	0.9089	0.7734
Scientific score	da	0.7521	0.8990	0.6872
GPQA main	fr	0.2789	0.3016	0.3039
GPQA main	da	0.2676	0.3016	0.2948
GPQA diamond	fr	0.3005	0.3212	0.3005
GPQA diamond	da	0.3057	0.2850	0.3005
SciFact binary	fr	0.4639	0.5090	0.4759
SciFact binary	da	0.7289	0.6898	0.7078
SciFact three-way	fr	0.5000	0.4699	0.4669
SciFact three-way	da	0.5151	0.4066	0.4096

Table 6: The individual head-seed accuracies underlying Table 2. All seeds share a task's encoded evaluation file. The adapted rows share one backbone adapter.

B Retrospective Selective Diagnostic

The diagnostic can be optimistic even when implemented correctly: it asks what threshold could have been chosen with access to the evaluation outcomes. A high value on synthetic references also inherits their validity limitations. No operating point from Table 7 is recommended for deployment.

C Excluded Historical Evidence

Earlier development explored additional architectures, larger QA pools, full-backbone fine-tuning, candidate families, and external baselines. We do not use those screens to support the principal quantitative claims here:

Evaluation	Frozen	Adapted
Scientific choice	1.000 ± 0.000	1.000 ± 0.000
Scientific verification	0.930 ± 0.013	0.905 ± 0.014
Scientific score	0.587 ± 0.359	0.496 ± 0.322
GPQA main	0.004 ± 0.001	0.003 ± 0.001
GPQA diamond	0.002 ± 0.003	0.000 ± 0.000
SciFact binary	0.256 ± 0.006	0.162 ± 0.021
SciFact three-way	0.012 ± 0.008	0.024 ± 0.037

Table 7: Oracle coverage at at most 5% observed error, mean ± sample SD across head seeds. Thresholds are selected using the evaluation labels and include complete confidence ties. These are *not* test results for a policy chosen on independent dev data and do not provide a deployment guarantee.

- The older `elite-v1` layout silently removed the question or instruction tail from many long-context rows, sometimes making identical inputs carry conflicting labels. Its architecture and data-scaling comparisons are not clean evidence of question-conditioned scientific improvements.
- The legacy `c/dag/da` candidate screen changed head optimization as well as adaptation. It is not the matched comparison in this manuscript.
- A non-frozen HF training path saved updated heads without the trained backbone. Its reloaded model cannot establish the effect of full fine-tuning or catastrophic forgetting.
- A RoBERTa verification baseline placed the claim after the abstract and silently truncated it in 52 of 340 validation pairs, with three more partially truncated. Its result is not a clean encoder-capacity comparison.
- Other generative and encoder diagnostics use different candidate families, cohorts, or metric conventions. Incomplete newer prediction and QA runs are outside the fixed, complete study selected here.

These exclusions are based on input validity, model identity, and protocol comparability, not on whether an experiment improves the headline score. The complete-passage matched study and its controls are not the defective `elite-v1` artifacts.

References

[1] S. Nie, F. Zhu, Z. You, et al. Large Language Diffusion Models. arXiv:2502.09992, 2025.

[2] TypeSafe AI. System One and Jev: typed decision models. Software and API documentation, 2026. `https://typesafe.ai`.

[3] Laya contributors. Laya: non-autoregressive typed decision models. Software, 2026. `https://github.com/NandhaKishorM/laya`.

[4] cbjev contributors. cbjev: typed decisions from one encoder pass. Software, 2026. `https://github.com/tomek7667/cbjev`.

[5] Kev contributors. Kev: open decision models on Qwen-family backbones. Software, 2026. `https://github.com/jaredpalmer/kev`.

[6] OpenJev contributors. OpenJev: an open System-One decision server. Software, 2026. `https://github.com/razorback16/openjev`.

[7] J. Austin, D. D. Johnson, J. Ho, D. Tarlow, and R. van den Berg. Structured Denoising Diffusion Models in Discrete State-Spaces. *NeurIPS*, 2021. arXiv:2107.03006.

[8] X. L. Li, J. Thickstun, I. Gulrajani, P. Liang, and T. B. Hashimoto. Diffusion-LM Improves Controllable Text Generation. *NeurIPS*, 2022. arXiv:2205.14217.

[9] Z. He, T. Sun, Q. Tang, K. Wang, X. Huang, and X. Qiu. DiffusionBERT: Improving Generative Masked Language Models with Diffusion Models. *ACL*, pp. 4521–4534, 2023. doi:10.18653/v1/2023.acl-long.248.

[10] S. S. Sahoo, M. Arriola, Y. Schiff, et al. Simple and Effective Masked Diffusion Language Models. *NeurIPS*, 2024. arXiv:2406.07524.

[11] A. Lou, C. Meng, and S. Ermon. Discrete Diffusion Modeling by Estimating the Ratios of the Data Distribution. *ICML*, 2024. arXiv:2310.16834.

[12] J. Ye, Z. Xie, L. Zheng, J. Gao, et al. Dream 7B: Diffusion Large Language Models. arXiv:2508.15487, 2025.

[13] G. Alain and Y. Bengio. Understanding Intermediate Layers Using Linear Classifier Probes. arXiv:1610.01644, 2016.

[14] S. Gururangan, A. Marasović, S. Swayamdipta, et al. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *ACL*, 2020. arXiv:2004.10964.

[15] E. J. Hu, Y. Shen, P. Wallis, et al. LoRA: Low-Rank Adaptation of Large Language Models. *ICLR*, 2022. arXiv:2106.09685.

[16] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On Calibration of Modern Neural Networks. *ICML*, 2017. arXiv:1706.04599.

[17] Z. Jiang, J. Araki, H. Ding, and G. Neubig. How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering. *TACL*, 2021. arXiv:2012.00955.

[18] Y. Geifman and R. El-Yaniv. SelectiveNet: A Deep Neural Network with an Integrated Reject Option. *ICML*, 2019. arXiv:1901.09192.

[19] R. El-Yaniv and Y. Wiener. On the Foundations of Noise-free Selective Classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

[20] A. N. Angelopoulos and S. Bates. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. arXiv:2107.07511, 2021.

[21] A. Poliak, J. Naradowsky, A. Haldar, R. Rudinger, and B. Van Durme. Hypothesis Only Baselines in Natural Language Inference. **SEM*, 2018. arXiv:1805.01042.

[22] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. R. Bowman, and N. A. Smith. Annotation Artifacts in Natural Language Inference Data. *NAACL*, 2018. arXiv:1803.02324.

[23] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, and H. Hajishirzi. Fact or Fiction: Verifying Scientific Claims. *EMNLP*, pp. 7534–7550, 2020. doi:10.18653/v1/2020.emnlp-main.609.

[24] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal. FEVER: a Large-scale Dataset for Fact Extraction and VERification. *NAACL*, 2018. arXiv:1803.05355.

[25] D. Rein, B. L. Hou, A. C. Stickland, et al. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *COLM*, 2024. arXiv:2311.12022.

[26] L. Zheng, W.-L. Chiang, Y. Sheng, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *NeurIPS*, 2023. arXiv:2306.05685.

[27] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. *EMNLP*, 2023. arXiv:2303.16634.

[28] C. Zheng, H. Zhou, F. Meng, J. Zhou, and M. Huang. Large Language Models Are Not Robust Multiple Choice Selectors. *ICLR*, 2024. arXiv:2309.03882.

[29] W. Cao, V. Mirjalili, and S. Raschka. Rank Consistent Ordinal Regression for Neural Networks with Application to Age Estimation. *Pattern Recognition Letters*, 140:325–331, 2020.

[30] A. L. Robles-Fernández. Sciev, version 0.2.2. Software and decision-head release, 2026. `https://github.com/alrobles/sciev/releases/tag/v0.2.2`.