EGRD: ASSESSING RESEARCH NOVELTY THROUGH EVIDENCE-GROUNDED RESIDUAL DECOMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Assessing the novelty of a research idea requires identifying what remains beyond prior work, even when coverage is dispersed across several papers or the contribution lies in a new combination of familiar components. We propose Evidence-Grounded Residual Decomposition (EGRD), a supervised framework for idea-level novelty prediction, assigning ordered grades relative to an explicit set of prior works published before the target idea. Given temporally admissible prior evidence, EGRD first estimates its joint coverage of the target idea, providing an evidence basis for subsequent novelty assessment. It then decomposes the target idea into latent semantic units and learns a support matrix aligning each unit with each prior. Aggregating support across priors yields unit residuals for insufficiently covered content, while within-prior co-support yields interaction residuals for potentially novel combinations. The model predicts novelty from these residuals together with coverage and contextual representations. To support training and evaluation, we construct IDEADELTA, which pairs research ideas with temporally admissible priors and provides coverage and novelty annotations across graph machine learning, computer vision, and time series. On temporally held-out ideas, EGRD outperforms the evaluated baselines across these domains and shows strong transfer performance on RINoBench. Ablations show that predictions depend on matched prior evidence and benefit from aggregating support across priors and representing residual contributions.

1 INTRODUCTION

Language models now propose research ideas faster than anyone can vet them, and deciding which of those ideas is new is the step that does not scale. Novelty is not a monotone function of textual similarity to any single prior work. We therefore study *evidence-conditioned* novelty assessment: grading a target idea against explicit prior works published before it, making judgments inspectable and enabling evidence interventions.

Most prior work takes a paper as the unit of assessment, from citation-pattern measures (Uzzi et al., 2013) to document-level embeddings (Cohan et al., 2020; Ostendorff et al., 2022) to systems that review a submission as a whole (Zhu et al., 2025). But a paper bundles contributions of uneven novelty, and one score over the bundle does not say which contribution it describes. Systems that do operate on ideas assemble the judgment as an LLM workflow: retrieve related work, then prompt an LLM to compare the target idea against it (Wang et al., 2024; da Silva et al., 2025; Afzal et al., 2026). Here the evidence is selected by the system being evaluated, so no verdict can be attributed to the retrieval or to the comparison, and the comparison itself is elicited by a prompt rather than learned against supervision. Evaluating this arrangement against expert judgments, Schopf & Färber (2026) find that the reasoning such models produce mirrors human rationales, yet that alignment does not translate into accurate grades.

Even when the target idea and the available prior works are held fixed, novelty assessment depends on how their relationship is represented. As illustrated in Figure 1(a), a prompt-based judge compares the target with earlier publications, but leaves the correspondence between individual parts of the target and their supporting evidence implicit. This correspondence matters for three reasons. First, support can be *dispersed* across multiple papers, each accounting for a different part of the idea. Second, similar overall coverage can leave substantively different content unsupported, such

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.

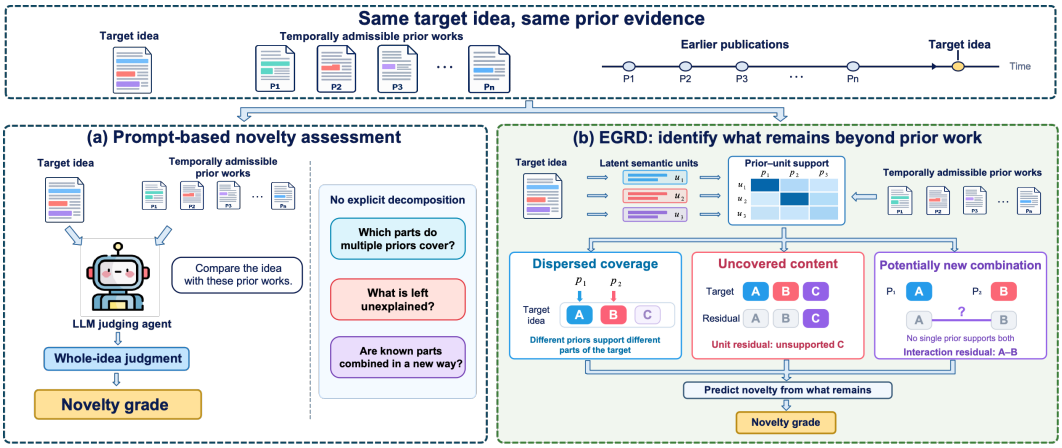

Figure 1: Contrasting novelty assessment approaches under the same target idea and temporally admissible prior works. **(a)** Prompt-based assessment leaves the correspondence between target content and supporting evidence implicit; **(b)** EGRD explicitly models this correspondence to aggregate dispersed support and represent unit and interaction residuals for novelty prediction.

as an implementation detail in one case and a central mechanism in another. Third, support for individual components does not necessarily establish support for their *combination*. These distinctions motivate an explicit representation of what the supplied evidence supports and what remains unsupported.

We propose Evidence-Grounded Residual Decomposition (EGRD) for idea-level novelty prediction. As illustrated in Figure 1(b), EGRD aligns latent semantic units with supplied prior works and combines residuals for unsupported content and potentially novel combinations with the evidence context to predict an ordered novelty grade.

Because existing benchmarks do not pair ideas with the evidence admissible when they were proposed, we also construct IDEADELTA to supply that pairing. Our contributions are:

- **IDEADELTA, a multi-domain evidence-paired dataset.** 21,111 ideas paired with temporally admissible priors and annotated at the idea and pair levels, with 92,876 pairs across graph machine learning, computer vision, and time series. Supplied rather than retrieved, the evidence supports controlled interventions that isolate the effect of prior work on novelty judgments.
- **Unit-level coverage over dispersed evidence.** We model coverage at the level of latent semantic units, aligning each unit with each prior through a learned support matrix. Noisy-OR aggregation lets several partially relevant papers jointly cover a unit none covers alone.
- **Typed residuals.** We split what the priors leave over into unit residuals for content no prior supports and interaction residuals for combinations no single prior puts together. Both are derived from the same support matrix.
- **Consistent gains across domains.** EGRD consistently outperforms all evaluated baselines across all three domains. We further evaluate transfer to expert annotations on RINoBench (Schopf & Färber, 2026) using a frozen checkpoint.

Model and training code for all three stages is available for review in an anonymized repository at `https://anonymous.4open.science/r/EGRD_model-86D7`, and IDEADELTA is available under an anonymized account at `https://huggingface.co/datasets/holaholaIDEADELTA/IdeaDelta`.

Related work is discussed in Appendix A.

2 DATASET

Novelty depends on an idea's relationship to prior work. IDEADELTA therefore provides three elements per example: a target idea, prior ideas available when it was proposed, and the contribution remaining beyond them.

2.1 TASK AND LABELS

The unit of annotation is a research *idea*: a single innovation claim extracted from a paper. A paper that introduces both a new objective and a new architecture therefore contributes two ideas, annotated separately. For a target idea x published at time t_x, the admissible priors are $\mathcal{P}_x = \{p_k \mid t_{p_k} < t_x\}$. This temporal constraint does real work: an idea published after the target says nothing about the target's novelty, however similar the two are.

IDEADELTA annotates each target at two levels. At the *pair* level, every target–prior pair receives one of four coverage levels: NOT COVERING, RELATED BUT NOT COVERING, PARTIAL COVER, or LARGE COVER. These record how much of the target that single prior accounts for. Each target's retained priors are ranked. At the *idea* level, the target receives an ordered novelty grade from LOW to HIGH. The two levels answer different questions: pair labels say what each piece of evidence covers, the idea grade says what survives all of it. A target can hold no strongly covering prior yet still score low, when several partial covers together leave little behind.

The idea-level grade is not an overall impression layered on top of the pair labels. We instead annotate four quantities per target and combine them by the fixed rule $\mathrm{raw}(x) = (1 - C_x) \cdot \Delta_x \cdot E_x \cdot (1 - \lambda S_x)$, where C_x is how much of the target the retained priors jointly cover, Δ_x the substantive contribution that remains, E_x whether the retrieved evidence suffices for a judgment, and S_x how much of the claimed difference is merely surface-level. Binning $\mathrm{raw}(x)$ yields the four ordered grades. The rule lowers novelty for exactly four reasons, each separately inspectable: prior work already covers the target, the remaining difference is not substantive, the evidence is too thin to judge, or the difference amounts to renaming.

2.2 CONSTRUCTION

Paper selection and idea extraction. We collect arXiv papers in graph machine learning, computer vision, and time series, preferring work that was subsequently published and, among that, work at major venues. An LLM extracts innovation claims from each title, abstract, and introduction. A second pass discards statements that make no substantive research claim. Appendix B.1 details the exclusion criteria.

Prior retrieval. For each target, nearest-neighbor search over idea embeddings (Cohan et al., 2020; Karpukhin et al., 2020) returns candidate priors published earlier, and an LLM adjudicates the uncertain similarity band. A cross-type retrieval pass links a target to relevant prior work of a different contribution type, which same-type retrieval alone would record as isolated. The retained priors are ranked for each target, and every pair stays within one research domain.

Annotation. A single judge (Zheng et al., 2023; Li et al., 2025), `deepseek-v4-pro`, annotates all stages. Pair coverage follows a hybrid rule: clearly covering and clearly unrelated pairs take their label directly, while the uncertain band goes to the judge, which reads the two ideas' technical content. The judge then estimates, for each target, joint coverage, residual contribution, evidence sufficiency, and surface-level difference, and the novelty rule above converts these into the idea-level grade. The labels are model-generated; we inspected samples from every stage. Appendix B.1 gives the full pipeline.

The result is $21{,}111$ annotated ideas and $92{,}876$ target–prior pairs spanning 2019-01 to 2026-09, split by date rather than at random so that no training target postdates an evaluation target; Appendix B.1 reports the composition per domain. Because the labels are model-generated, we also check that they respond to the retrieved evidence rather than to target metadata alone, and that novelty decreases when a prior already covers the target contribution. Appendix B.2 reports these checks.

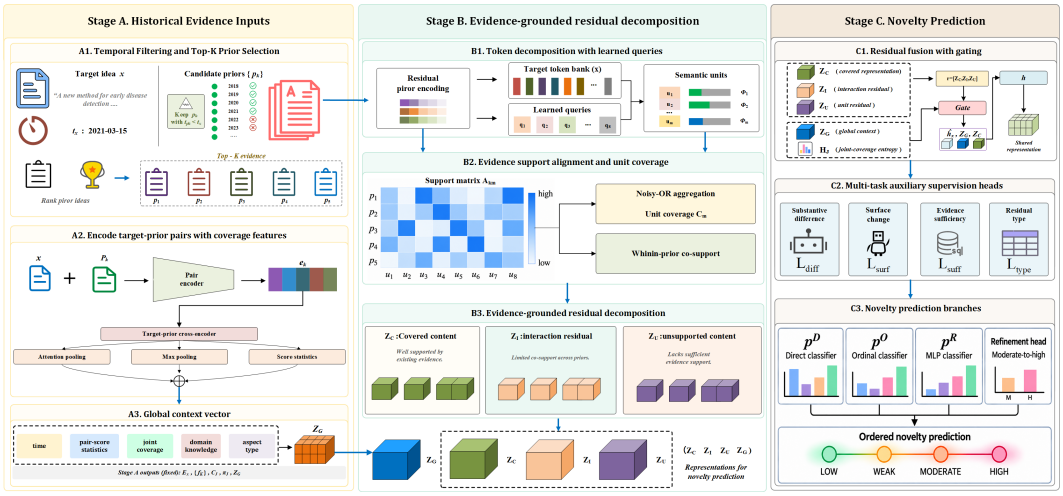

Figure 2: Overview of Evidence-Grounded Residual Decomposition (EGRD). **(A)** Historical evidence selection and coverage estimation provide fixed inputs for novelty assessment. **(B)** Learned queries decompose the target into semantic units, whose support from prior ideas yields covered content, unit residuals for content with limited support, and interaction residuals for combinations with limited co-support within any single prior. **(C)** The residuals are fused with gated global context and combined with covered content and the target representation for novelty prediction, using auxiliary supervision, three prediction branches, and a moderate-to-high refinement.

3 METHODOLOGY

Given a target research idea and a set of ideas proposed before it, EGRD assesses the target's novelty relative to the supplied evidence in three stages (Figure 2). Stage A selects temporally admissible priors and estimates how much of the target they already account for. Stage B decomposes the target against that evidence into two complementary components: unit residuals capture semantic content with insufficient historical support, and interaction residuals capture potentially novel combinations of those units. Stage C predicts an ordered novelty grade from LOW to HIGH from the residual representations. Throughout, $[\cdot;\cdot]$ denotes concatenation, $\odot$ elementwise multiplication, d the encoder dimension, σ the sigmoid, ε a small constant added for numerical stability, and $\mathcal{V}$ the non-padding positions of the target; padded priors are masked out.

3.1 STAGE A: HISTORICAL EVIDENCE INPUTS

Given a target idea x with timestamp t_x, we restrict candidate priors to those predating it, $\mathcal{P}_x = \{p_k \mid t_{p_k} < t_x\}$, and retain the top-$K$ ranked candidates as the evidence set $\mathcal{E}_x \subseteq \mathcal{P}_x$. We use $K = 5$ in our experiments and retain all candidates when fewer than K are available. Each retained prior p_k carries a vector of pair coverage features $\boldsymbol{f}_k \in \mathbb{R}^{11}$.

To estimate coverage across the evidence set, a cross-encoder (Nogueira & Cho, 2019) encodes each target–prior pair and a multiple-instance model (Ilse et al., 2018) aggregates the pair representations through attention pooling and maximum pooling. Concatenated with evidence-score statistics, the pooled representations predict joint coverage C_J and a four-class coverage distribution $\boldsymbol{\pi}^J$. The retained evidence, pair features, and coverage predictions serve as fixed inputs to Stages B and C, grounding the residual decomposition and providing coverage information for novelty prediction.

The global context $\boldsymbol{z}_G$ projects twenty-one scalars—the target timestamp, ten pair-score statistics, and ten joint-coverage features—together with embeddings of the target's aspect and contribution type. Appendix C.2 specifies these features, the pair-scoring rule, and the joint model.

3.2 STAGE B: EVIDENCE-GROUNDED RESIDUAL DECOMPOSITION

Stage B decomposes the target into semantic units and estimates their support from the retained evidence, representing the unexplained contribution through unit residuals for content with limited

historical support and interaction residuals for combinations with limited joint support from any single prior.

Semantic decomposition. A residual cross-encoder, initialized from the Stage A model and trained separately from it, produces a pooled representation $\boldsymbol{v}_k$ for each target–prior pair, which we augment with projected pair coverage features as $\widetilde{\boldsymbol{v}}_k = \boldsymbol{v}_k + f_P(\boldsymbol{f}_k)$. Averaging the contextual token states at matching positions across valid pairs yields a shared token bank $\{\boldsymbol{h}_j\}_{j=1}^{N}$ over the N target tokens. We then use M learned queries $\boldsymbol{q}_m$ to softly assign these tokens to semantic units (Carion et al., 2020; Locatello et al., 2020). We use $M = 8$ in our experiments. The queries compete for each token, $B_{mj} = \mathrm{softmax}_m(\boldsymbol{q}_m^\top \boldsymbol{h}_j/\sqrt{d})$, and the resulting assignments are normalized over tokens within each unit, $\alpha_{mj} = B_{mj}/(\sum_{j' \in \mathcal{V}} B_{mj'} + \varepsilon)$, giving $\boldsymbol{u}_m = \sum_{j \in \mathcal{V}} \alpha_{mj} \boldsymbol{h}_j$. The units have no predefined semantic roles. A learned activity score $\phi_m = \sigma(\boldsymbol{w}_\phi^\top \boldsymbol{u}_m + b_\phi)$ controls each unit's contribution, while an orthogonality penalty on the attention matrix discourages redundant token assignments.

Evidence-aligned unit residuals. A learned support matrix estimates how strongly each prior p_k supports each unit $\boldsymbol{u}_m$, $A_{km} = \sigma(f_A([\boldsymbol{u}_m; \widetilde{\boldsymbol{v}}_k; \boldsymbol{u}_m \odot \widetilde{\boldsymbol{v}}_k; |\boldsymbol{u}_m - \widetilde{\boldsymbol{v}}_k|; \boldsymbol{f}_k]))$, where f_A is an MLP. Support is estimated separately for every target–prior pair, so different priors can account for different units. Over the $n = |\mathcal{E}_x|$ retained priors, noisy-OR (Pearl, 1988) aggregates support into unit coverage $c_m = 1 - \prod_{k=1}^{n}(1 - A_{km})$; its complement, weighted by unit activity, defines the residual weight $r_m = \phi_m(1 - c_m)$ and the unit residual $\boldsymbol{z}_U = M^{-1} \sum_m r_m \boldsymbol{u}_m$. Noisy-OR lets several partially relevant priors jointly cover a unit that none of them covers alone; we use it as a differentiable aggregation rule, not as a claim that the priors are independent sources of support (Appendix C.3). An active unit with little historical support therefore receives a large residual weight.

Dividing by the fixed unit count M preserves residual magnitude. We pool covered content similarly as $\boldsymbol{z}_C = M^{-1} \sum_m \phi_m c_m \boldsymbol{u}_m$. A learned, activity-weighted pooling of c_m reconstructs joint coverage $\widehat{C}_U$, supervised by annotated coverage and the fixed Stage A prediction, without unit-level alignment labels.

Interaction residuals. Units that are individually covered may still form a new combination. For each unit pair $i < j$, an MLP maps their concatenation, elementwise product, and absolute difference to an interaction representation $\boldsymbol{t}_{ij}$, and a sigmoid head predicts its strength g_{ij}. We measure historical co-support by the largest product of the two units' support strengths within a single prior, $\bar{A}_{ij} = \max_k A_{ki} A_{kj}$, which gives the interaction residual weight $r_{ij} = g_{ij}(1 - \bar{A}_{ij})$ and the pooled representation $\boldsymbol{z}_I = \sum_{i<j} \eta_{ij} r_{ij} \boldsymbol{t}_{ij}$, where the learned attention η_{ij} includes a $\log(r_{ij} + \varepsilon)$ bias that favors pairs with larger residual weights. When different priors support the two units but no single prior strongly supports both, their unit residuals can be small while their interaction residual remains large, provided that g_{ij} is high. Co-support is only a proxy for evidence that the combination already exists: a prior that supports both units need not cover the specific way the target combines them. Multiplication by r_{ij} after attention normalization preserves residual magnitude. An auxiliary classifier on $\boldsymbol{z}_I$ is trained with idea-level labels for nontrivial combinations (Appendix C.3).

3.3 Stage C: Novelty Prediction

Stage C combines residuals with evidence context to predict a novelty grade. A learned gate γ, conditioned on the residuals, context, and coverage uncertainty, yields $\boldsymbol{z}_R = f_R([\boldsymbol{z}_U; \boldsymbol{z}_I; \gamma \boldsymbol{z}_G])$. We combine this with covered content, the target representation $\hbar_x$, and ungated context: $\boldsymbol{h} = \mathrm{LayerNorm}([\boldsymbol{z}_R; \boldsymbol{z}_C; \hbar_x; \boldsymbol{z}_G])$.

Auxiliary heads predict substantive difference, surface change, evidence sufficiency, and residual type. These predictions, together with $\boldsymbol{h}$ and joint coverage, feed three novelty branches: a direct linear classifier, a cumulative ordinal head that models grade order (Cao et al., 2020), and a two-layer MLP classifier. Each branch produces a four-class distribution, and their convex mixture determines an initial grade.

MODERATE and HIGH account for most of the confusion between grades, so a moderate-to-high refinement revisits that boundary once the initial grade is set. The refinement upgrades an initial

Table 1: Baselines and our model on the IdeaDelta evaluation set, 3,352 ideas, split by domain. A majority-class predictor scores 0.312 Accuracy on Graph ML, 0.356 on computer vision and 0.328 on time series, with QWK 0 by construction.

		Graph ML				Computer vision				Time series			
System	Backbone	Accuracy	Macro-F1	QWK	MAE ↓	Accuracy	Macro-F1	QWK	MAE ↓	Accuracy	Macro-F1	QWK	MAE ↓
Dense RAG, 5 priors	Qwen3.5	0.360	0.217	0.129	0.867	0.367	0.248	0.191	0.801	0.318	0.248	0.190	0.954
	deepseek-flash	0.371	0.236	0.152	0.842	0.369	0.232	0.154	0.801	0.326	0.224	0.205	0.919
	gpt-5.6-luna	0.362	0.262	0.253	0.810	0.372	0.289	0.267	0.787	0.346	0.295	0.292	0.883
Direct, 5 priors	Qwen3.5	0.300	0.266	0.319	0.925	0.308	0.264	0.300	0.876	0.291	0.283	0.327	0.956
	deepseek-flash	0.438	0.444	0.558	0.706	0.409	0.418	0.521	0.728	0.416	0.415	0.560	0.716
	gpt-5.6-luna	0.432	0.439	0.579	0.657	0.420	0.430	0.561	0.661	0.445	0.436	0.576	0.653
Schopf 2-stage	Qwen3.5	0.324	0.286	0.335	0.831	0.346	0.293	0.338	0.777	0.344	0.333	0.413	0.793
	deepseek-flash	0.443	0.458	0.560	0.639	0.416	0.411	0.501	0.676	0.442	0.447	0.553	0.654
	gpt-5.6-luna	0.391	0.377	0.487	0.698	0.393	0.364	0.458	0.683	0.392	0.379	0.512	0.688
UKPLab 3-stage	Qwen3.5	0.280	0.161	0.125	0.928	0.289	0.150	0.106	0.873	0.274	0.180	0.132	0.973
	deepseek-flash	0.351	0.267	0.301	0.761	0.376	0.272	0.259	0.741	0.386	0.289	0.290	0.727
	gpt-5.6-luna	0.318	0.226	0.266	0.800	0.367	0.248	0.304	0.721	0.365	0.267	0.296	0.743
GraphMind	Qwen3.5	0.264	0.205	0.154	1.027	0.264	0.182	0.108	0.959	0.244	0.204	0.155	1.051
	deepseek-flash	0.291	0.227	−0.007	1.095	0.342	0.266	0.154	0.971	0.365	0.249	0.159	0.916
	gpt-5.6-luna	0.289	0.255	0.046	1.003	0.381	0.328	0.304	0.755	0.380	0.327	0.305	0.750
DeepReviewer	DeepReviewer-7B	0.294	0.237	0.034	0.965	0.325	0.235	−0.043	0.972	0.296	0.222	−0.017	0.964
DeBERTa, flat 5 priors	DeBERTa-v3-base	0.552	0.551	0.722	0.499	0.528	0.535	0.680	0.529	0.598	0.584	0.733	0.452
EGRD (ours)	—	**0.799**	**0.801**	**0.909**	**0.201**	**0.761**	**0.767**	**0.882**	**0.242**	**0.789**	**0.776**	**0.898**	**0.215**

MODERATE prediction when a weighted combination of the direct and ordinal HIGH-class probabilities and a separate binary-head score exceeds a threshold.

Training. We initialize the residual encoder from the Stage A joint-coverage checkpoint and jointly optimize it with the decomposition and prediction modules, while keeping the Stage A models fixed. The objective combines classification of the novelty grade and the residual type, smooth-L_1 supervision of five annotated scalars, and regularization of the decomposition against redundant token assignments, unused units, and coinciding unit coverages. The five scalars are substantive difference, surface change, evidence sufficiency, the reconstructed coverage $\widehat{C}_U$, and a continuous novelty score. Ranking supervision complements scalar regression by encouraging predictions to preserve the annotated order; the coverage term additionally penalizes disagreement with the fixed Stage A prediction, weighted down where that prediction has higher entropy. Appendix C.5 states the full objective, defines each term, and reports the coefficients.

4 EXPERIMENTS

The evaluation asks three things: whether EGRD improves on prompted baselines and a supervised encoder baseline (§4.1), whether what it learns survives labels it was not trained on (§4.2), and what its predictions depend on (§4.3, §4.4). Supplied evidence enables these interventions: all systems except dense RAG read the same five admissible priors per test target. All rows report Accuracy, Macro-F1, QWK and MAE; Appendix D.1 gives the splits, hyperparameters and scoring code.

4.1 MAIN RESULTS

Table 1 compares a *direct* single call; *Schopf 2-stage* (Schopf & Färber, 2026) and *UKPLab 3-stage* (Afzal et al., 2026) prompt chains; *GraphMind* (da Silva et al., 2025), which labels priors as supporting or contrasting before rating; and *dense RAG*, which uses the direct prompt with five priors retrieved by dense similarity. These workflows use three LLM backbones. We also evaluate *DeepReviewer-7B* (Zhu et al., 2025), fine-tuned on paper review, and a supervised *DeBERTa* encoder baseline with five priors.

The effect of the Schopf chain declines monotonically with how well the backbone scores without it. Against a single five-prior call, the two-stage chain gains in all three domains on Qwen3.5, the weakest backbone here, on both Accuracy and QWK; on deepseek-flash the QWK gain is already gone; and on gpt-5.6-luna, whose unaided QWK is the highest of the three, it loses in all three domains on both metrics. We interpret this pattern as intermediate evidence interpretations helping weaker backbones while potentially interfering with stronger backbones' own judgments. UKPLab 3-stage and GraphMind also fall below the direct call on both metrics in all nine backbone–domain

Table 2: Results on RINoBench with expert ratings mapped to four ordinal grades.

System	Accuracy	Macro-F1	QWK	MAE ↓
Same Qwen3.5 backbone, zero-shot				
Qwen3.5	0.3574	0.2244	0.1451	0.7726
GraphMind	0.3357	0.2281	0.1469	0.7653
UKPLab 3-stage	0.3466	0.1892	0.0808	**0.7220**
Released predictions of the benchmark authors, folded to four levels				
GPT-5	0.3646	0.2457	0.0955	0.8051
o3	0.3971	0.2370	0.0606	0.8087
gpt-oss-120b	0.4152	0.2249	0.0392	0.8375
Llama-3.1-8B-Instruct	0.3755	0.1961	−0.0011	0.8773
DeepSeek-R1	**0.4224**	0.1873	0.0283	0.8700
Llama-4-Scout-17B-16E	0.4116	0.1864	0.0215	0.8809
Llama-3.3-70B-Instruct	0.4188	0.1531	0.0051	0.9025
Reviewer model fine-tuned on paper review, local weights				
DeepReviewer-7B	0.1913	0.1425	0.0145	1.3610
EGRD (ours)	0.4152	**0.3407**	**0.2245**	0.8087

combinations. Additional structure alone therefore does not guarantee gains: the effect of prompt chaining depends on both the backbone and the specific workflow.

Replacing the supplied priors with query-time dense retrieval costs ordinal agreement on every backbone: against the direct call on the same backbone and domain, Macro-F1 and QWK fall in all nine cells. Accuracy moves the other way on the weakest backbone, rising on Qwen3.5 while Macro-F1 falls. Accuracy in these rows is 0.318–0.372, at or just above the majority-class band. The group also compresses the backbone differences Table 1 otherwise shows, its Accuracy spanning at most a quarter as much in any domain as the direct call's. What a system is given to compare against therefore moves the judgment more than which model or chain performs it. That sharpens the evidence check of Appendix B.2: random priors there reduce label prediction to near-chance, establishing only that the evidence must be on topic, whereas dense retrieval is on topic and the gap persists. That persistence shows the candidate lists carry target–prior relationships generic similarity search does not recover, which is what the cross-type retrieval pass and the adjudicated similarity band of IDEADELTA are for.

The supervised DeBERTa encoder is the strongest baseline on all four metrics in all three domains, indicating that IDEADELTA provides useful training signal for a general-purpose encoder. EGRD further improves Accuracy by 24.7, 23.3 and 19.1 percentage points on Graph ML, computer vision and time series, respectively, and improves all other metrics. Compared with direct prompting and flat text encoding, EGRD explicitly models evidence coverage and residual contributions, which may help distinguish substantive novelty from semantic differences. DeepReviewer-7B has near-zero QWK in all three domains, negative in two: paper-review fine-tuning alone does not recover the grade ordering here.

4.2 TRANSFER EVALUATION

To evaluate transfer beyond IDEADELTA's model-generated labels, we test the frozen checkpoint from Table 1 on RINoBench (Schopf & Färber, 2026), which contains 277 expert-rated ideas spanning machine learning. Table 2 reports agreement with expert ratings mapped to four ordinal grades. Appendix D.2 details the data, label mapping, inference procedure, and baseline evaluation.

EGRD leads both metrics that reward ordering the ideas the way the experts did, and on QWK no other system reaches two thirds of its value while most of the table sits near zero. Accuracy and MAE belong to other rows, but reading those leads as closer agreement mistakes what the folded scale rewards. Merging the top two expert levels leaves 41.5% of the ideas in a single class, so answering with that class alone scores 0.4152 Accuracy and a competitive MAE while ordering nothing, and the systems that lead those two metrics come close to doing exactly that: their ordinal agreement is at or near zero, the value a constant answer earns by construction. Our row reaches the same 0.4152 from a prediction spread over all four grades, $16/36/76/149$ against a gold $15/60/87/115$, and it is that spread the rank-sensitive metrics register.

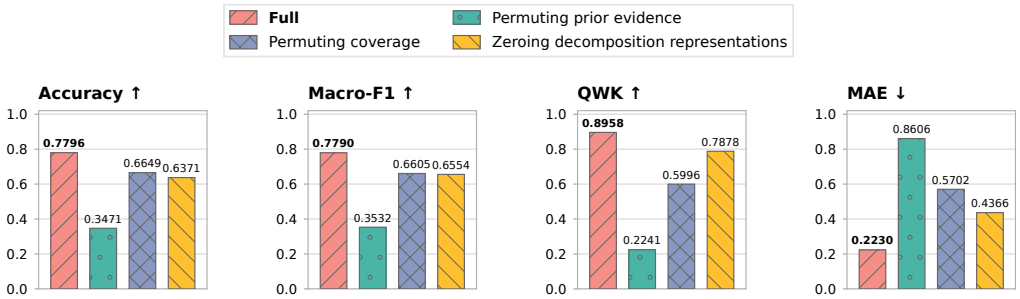

Figure 3: Pipeline-stage ablations on IdeaDelta. Full is the unablated checkpoint of Table 1; each intervention is applied at inference without retraining. Arrows indicate the preferred direction for each metric.

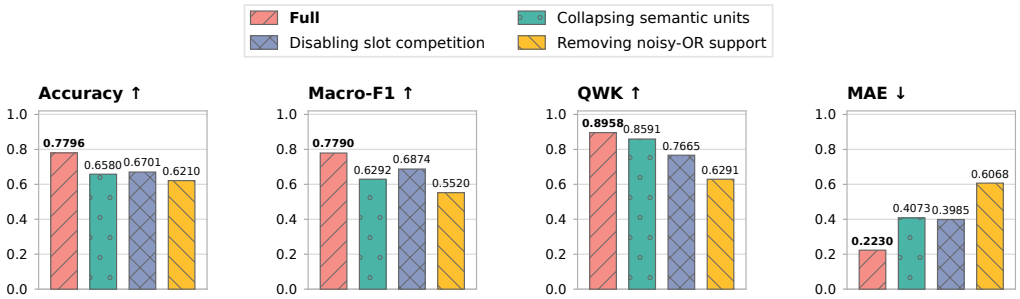

Figure 4: Decomposition-design ablations on IdeaDelta. Each variant is retrained without the corresponding component and compared with the same Full model. Arrows indicate the preferred direction for each metric.

Two properties of the transfer limit agreement. First, IDEADELTA covers three fields, whereas RINoBench spans machine learning as a whole: only 30 of its 277 test ideas fall within these three fields. The remaining 247 ideas therefore test transfer beyond the domains covered by our task-specific training data. This shift may limit the model's ability to parse target units and identify their support in prior work, making the decomposition less reliable. The second is the label scale: mapping five expert levels to four grades merges the top two, where many examples lie, so differences within that class cannot be evaluated. We therefore interpret the results as partial agreement between residual structure learned from model-generated labels and expert judgment.

4.3 ABLATION

Holding the target fixed, three evidence and representation interventions in Figure 3 test prediction dependencies at successive pipeline stages. **Permuting prior evidence** shuffles the retrieved prior texts across targets, breaking the correspondence between each target and its evidence in Stage A. **Permuting coverage** shuffles only the joint coverage score produced by Stage A while leaving every prior text in place. **Zeroing decomposition representations** sets the unit residual $\boldsymbol{z}_U$, interaction residual $\boldsymbol{z}_I$ and covered-content representation $\boldsymbol{z}_C$ to zero, removing these three outputs of Stage B together. All three reduce performance on every metric.

Permuting prior evidence causes the largest loss across all four metrics, establishing target–prior correspondence as a critical basis for the model's predictions. Permuting coverage preserves the prior texts but sharply reduces ordinal agreement and increases grade errors, showing that the remaining inputs cannot compensate for a corrupted coverage signal. Zeroing decomposition representations produces more incorrect predictions than permuting coverage, but smaller grade errors on average. Together, these results suggest complementary roles for coverage in maintaining ordinal agreement and decomposition representations in assigning the exact grade.

Figure 4 shows that the three mechanisms inside the decomposition are not interchangeable; each removal degrades a different part of the prediction. **Collapsing semantic units** from eight to one costs on Accuracy, Macro-F1 and MAE while QWK barely moves. **Disabling slot competition**,

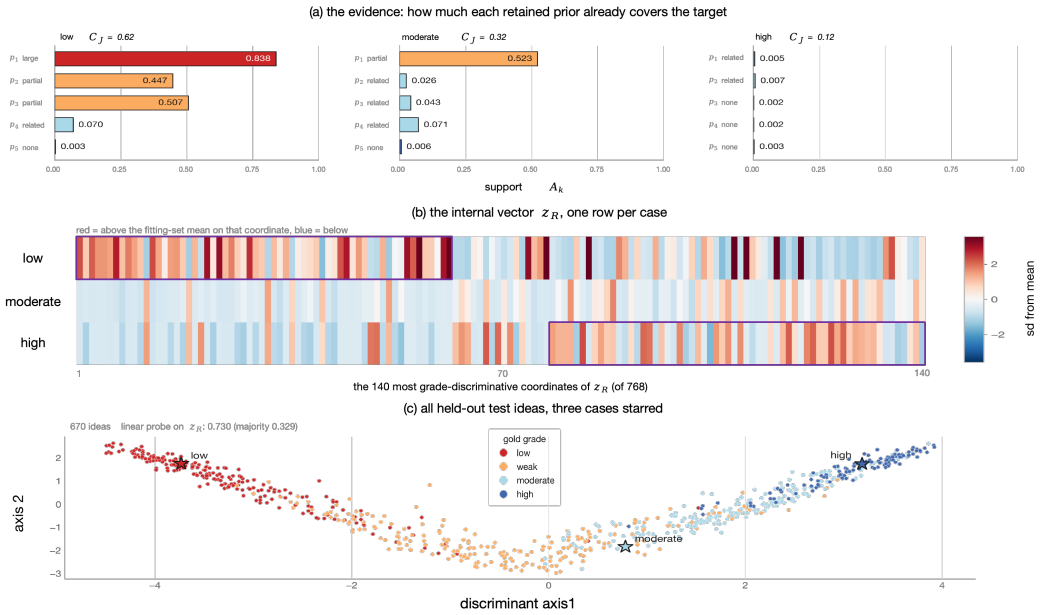

Figure 5: Three test targets, graded low, moderate and high, traced through the model. **(a)** The support A_k assigned to each retained prior, against the coverage label the annotation carries; C_J is the predicted joint coverage. **(b)** The residual representation $\boldsymbol{z}_R$ of each case, as deviations from the fitting-set mean. **(c)** $\boldsymbol{z}_R$ for every held-out test idea, coloured by gold grade, the three cases starred.

which removes the softmax through which the units compete for each token, reverses that pattern: QWK falls while MAE stays where collapsing the units left it. **Removing noisy-OR support** costs the most of the three on all four metrics and most of all on the rank-sensitive ones.

Appendix D.3 reports parameter sensitivity experiments around the default configuration.

4.4 Case Study

Figure 5 illustrates the relationship between evidence coverage, residual embeddings, and novelty through three correctly predicted cases. Joint coverage decreases from the low- to the high-novelty case, illustrating how higher novelty corresponds to less coverage by prior work in these examples. Their residual embeddings $\boldsymbol{z}_R$ show contrasting spatial patterns in the heatmap: the low-novelty case has predominantly above-mean values in the left portion, whereas the high-novelty case shows this pattern in the right portion; the moderate case has weaker deviations overall. This contrast shows how novelty differences are reflected in the displayed embedding dimensions.

The supervised discriminant projection reveals a broader pattern: held-out ideas are broadly ordered by novelty along the first axis, although adjacent grades overlap. The three cases lie in regions consistent with their grades, linking their individual embedding patterns to the overall distribution. Together, these observations connect the coverage differences in the examples to a representation space whose organization reflects novelty levels. Additional case-study details are provided in Appendix D.4.

5 Conclusion

EGRD grades an idea by what an explicit set of admissible priors leaves over, estimating coverage at the level of semantic units and predicting the grade from the residuals that survive alignment with the evidence. Supplying that evidence rather than retrieving it is what lets the design be checked: permuting the priors lowers Accuracy to the majority-class band, removing noisy-OR aggregation costs more on every metric than any other change to the decomposition, and a cross-validated probe finds grade information linearly accessible in the residual representation. EGRD outperforms all evaluated baselines. On expert labels, the ordering transfers while the level does not (App. E).

REFERENCES

Osama Mohammed Afzal, Preslav Nakov, Tom Hope, and Iryna Gurevych. Beyond "not novel enough": Enriching scholarly critique with LLM-assisted feedback. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2648–2671, Rabat, Morocco, 2026. Association for Computational Linguistics. URL https://aclanthology.org/2026.eacl-long.121/.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. ResearchAgent: Iterative research idea generation over scientific literature with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6709–6738, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.naacl-long.342/.

Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020. doi: 10.1016/j.patrec.2020.11.008.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020*, volume 12346 of *Lecture Notes in Computer Science*, pp. 213–229. Springer, 2020. doi: 10.1007/978-3-030-58452-8_13.

Yuan Chang, Ziyue Li, Hengyuan Zhang, Yuanbo Kong, Yanru Wu, Hayden Kwok-Hay So, Zhijiang Guo, Liya Zhu, and Ngai Wong. TreeReview: A dynamic tree of questions framework for deep and efficient LLM-based scientific peer review. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 15651–15682, Suzhou, China, 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.emnlp-main.790/.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2270–2282, Online, 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.acl-main.207/.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Jacob Cohen. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, 1968. doi: 10.1037/h0026256.

Italo Luis da Silva, Hanqi Yan, Lin Gui, and Yulan He. GraphMind: Interactive novelty assessment system for accelerating scientific discovery. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2025. URL https://arxiv.org/abs/2510.15706.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2111.09543.

Maximilian Ilse, Jakub M. 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.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. AgentReview: Exploring peer review dynamics with LLM agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1208–1226, Miami, Florida, USA, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.emnlp-main.70/.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online, 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.emnlp-main.550/.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2757–2791, Suzhou, China, 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.emnlp-main.138/.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, Daniel McFarland, and James Zou. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8), 2024. doi: 10.1056/AIoa2400196.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems 33*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/8511df98c02ab60aea1b2356c013bc0f-Abstract.html.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024. URL https://arxiv.org/abs/2408.06292.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 12076–12100, Singapore, 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.emnlp-main.741/.

Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*, 2019. URL https://arxiv.org/abs/1901.04085.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 11670–11688, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.802/.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using LLM agents as research assistants. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 5977–6043, Suzhou, China, 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.findings-emnlp.320/.

Tim Schopf and Michael Färber. Is this idea novel? an automated benchmark for judgment of research ideas. In *Proceedings of the 2026 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC)*, 2026. URL https://arxiv.org/abs/2603.10303.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ea94957d81b1c1caf87ef5319fa6b467-Abstract-Conference.html.

Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. Atypical combinations and scientific impact. *Science*, 342(6157):468–472, 2013. doi: 10.1126/science.1240474.

Juraj Vladika and Florian Matthes. Comparing knowledge sources for open-domain scientific claim verification. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2103–2114, St. Julian's, Malta, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.128/.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550, Online, 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.emnlp-main.609/.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 61–76, Seattle, United States, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-naacl.6/.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. SciMON: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 279–299, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.18/.

Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. A closer look at claim decomposition. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pp. 153–175, Mexico City, Mexico, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.starsem-1.13/.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. Long-form factuality in large language models. In *Advances in Neural Information Processing Systems 37*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/937ae0e83eb08d2cb8627fe1def8c751-Abstract-Conference.html.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 641–649, 2024. doi: 10.1145/3626772.3657878.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36, Datasets and Benchmarks Track*, 2023. URL https://arxiv.org/abs/2306.05685.

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. From automation to autonomy: A survey on large language models in scientific discovery. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 17733–17750, Suzhou, China, 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.emnlp-main.895/.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. DeepReview: Improving LLM-based paper review with human-like deep thinking process. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29330–29355, Vienna, Austria, 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.acl-long.1420/.

A Related Work

Automated research and idea generation. Language models are increasingly used to carry out parts of the research process itself. Wang et al. (2024) retrieve inspirations from past literature and optimise explicitly for novelty by comparing a candidate against prior papers; Baek et al. (2025) refine proposals against a set of LLM reviewing agents that give structured feedback along several dimensions; and Lu et al. (2024) and Schmidgall et al. (2025) close the loop from ideation through experiments to a written paper and a simulated review. Si et al. (2025) evaluate what such pipelines produce, finding the generated ideas rated more novel than human-written ones but less feasible; Zheng et al. (2025) survey the progression of these systems toward greater autonomy. Every one of these systems contains an internal novelty check, and it is the check rather than the generator that bounds them: once proposals are cheap, the output of the pipeline is whatever its filter admits. The filter is nonetheless the component these systems evaluate least directly, judged only through the ideas that eventually emerge downstream of it.

Judging the novelty of scientific work. Studied on its own, the check has passed through several generations of method. The earliest operate on citation statistics, measuring how unusual the combinations of references in a paper are against historical co-occurrence (Uzzi et al., 2013); they require the work to already exist and be cited, and reduce novelty to patterns of co-occurrence rather than to what a paper says. Document-level semantic representations replace surface statistics with learned relatedness, scoring how close a submission is to prior papers through citation-informed embeddings (Cohan et al., 2020; Ostendorff et al., 2022), but they compare documents as points and recover nothing about which content is shared. Most recent systems retrieve related literature and hand the comparison to a language model. Liang et al. (2024) obtain structured feedback from a single prompt, of which significance and novelty is one section, and report partial overlap with human reviews; Zhu et al. (2025) train a reviewer model on a multi-stage process of structured analysis, retrieval, and evidence-based argumentation; Chang et al. (2025) decompose the review into a tree of questions resolved from leaf to root, and Jin et al. (2024) simulate the review process itself to expose how much of the outcome is attributable to reviewer bias rather than to the paper; Afzal et al. (2026) split the judgment into claim extraction, retrieval and synthesis of related work, and structured comparison, motivating the design with an analysis of human-written novelty reviews; and da Silva et al. (2025) build an interactive system on the same retrieve-then-compare pattern. Schopf & Färber (2026) supply the evaluation counterpart and report that model ratings still diverge substantially from expert ones even when the accompanying reasoning appears sound. Two properties persist across these generations, and both are structural rather than incidental. The evidence is retrieved by the system under evaluation, so its verdicts are unattributable: an error cannot be assigned to the retrieval or to the judgment, and the two therefore cannot be improved, or even measured, separately. We remove this by conditioning on a given, temporally admissible evidence set, which additionally turns the evidence into a variable that can be intervened on. The second property is that the target is scored as an undivided whole, against the closest prior work or against a retrieved set read as a block. Such a comparison returns one number per target, and no post-processing of that number can recover which part of the idea it is about; the information is discarded before the score is formed, not lost afterwards.

Decomposed verification against evidence. Saying which part remains requires decomposing the target, and outside novelty assessment this has become the standard recipe for factuality evaluation: break a text into smaller propositions and check each against retrieved evidence. Min et al. (2023) decompose a passage into atomic facts, verify each against a knowledge source, and report the supported fraction; Wei et al. (2024) scale the same procedure with search as the evidence source; and in the scientific domain the corresponding task checks a stated claim against candidate abstracts while identifying the supporting rationales (Wadden et al., 2020; 2022), where the choice of evidence source itself moves the verdict (Vladika & Matthes, 2024). Three properties of the recipe do not survive the move to research ideas. The propositions are assumed to be given, whereas an idea carries no stated claims, and the decomposition step is not a neutral preprocessor: Wanner et al. (2024) show that scores of this family shift materially with the decomposition used. Support is treated as a label, whereas a prior work can cover part of a unit and not the rest. And each unit is verified independently before the results are averaged, so the score is a function of the per-unit coverages alone and cannot express whether a combination of units has precedent. We accordingly

Table 3: Composition of IDEADELTA: counts of ideas and of target–prior pairs, and the idea-level novelty distribution, by source domain. Pairs are formed within a source domain only, so *All* is the union of the three subsets rather than a fourth collection.

	Graph ML	Computer vision	Time series	All
Ideas	5,225	8,291	7,595	21,111
Pairs	21,183	36,238	35,455	92,876
LOW	819	901	1,724	3,444
WEAK	1,240	1,711	1,614	4,565
MODERATE	2,247	2,777	2,173	7,197
HIGH	919	2,902	2,084	5,905

induce the units rather than reading them off the text, measure support as a graded quantity, and add a residual defined over pairs of units.

B DATASET DETAILS

B.1 DATASET CONSTRUCTION PIPELINE

Sources and idea extraction. Papers are drawn from arXiv in three areas: graph machine learning, computer vision, and time series. Selection prefers work that was subsequently published and, among that, papers at major venues. From each paper's title, abstract and introduction an LLM extracts innovation claims and assigns each an innovation aspect; a second LLM pass discards claims that are not innovation claims at all, including open-source and reproducibility statements, bare state-of-the-art assertions, and engineering deliverables.

Candidate priors. Idea texts are embedded and indexed for top-K nearest-neighbour recall, and candidate relations are formed under a hybrid rule: high similarity is accepted directly, low similarity is rejected directly, and the uncertain band is passed to an LLM. The base build pairs ideas of the same type only, which would record an idea whose closest relative is of another type as isolated. A cross-type pass recovers these relations. The retained priors are ranked for each target. Pairing stays within a source domain, and no cross-domain pair is labelled, so the three single-domain subsets and their union address the same ideas under identifiers namespaced by domain. The assembled dataset passes its own validation checks with no duplicate pair keys, no duplicate novelty targets, no invalid pair labels, no target left without a pair, and no temporal violation.

Annotation. All annotation stages use the same judge, `deepseek-v4-pro`. It labels pair coverage over the four classes. Pairs in the ambiguous band are re-judged rather than accepted from the similarity rule. This band includes related but not covering, partial cover, large cover, and high-similarity pairs the graph left unconnected. It then estimates, for each target, the substantive residual that survives the retained evidence, the sufficiency of that evidence, and the extent to which the claimed difference is surface-level. Idea-level grades follow by applying the novelty rule of Sec. 2.1 to those quantities. Labels are model-generated; each stage was reviewed by sampled manual inspection.

Composition. IDEADELTA contains 21,111 annotated ideas and 92,876 annotated target–prior pairs, with target dates from 2019-01 to 2026-09 (Table 3). Each idea carries its source domain, and the *All* column is the union of the three domain subsets rather than a fourth collection.

Splits. Splits are date-ordered rather than random, so that no training target is dated after an evaluation target. The cuts fall at 2025-02 and 2025-11, giving $13{,}815/3{,}944/3{,}352$ ideas for training, validation, and test.

B.2 VALIDATION OF RESEARCH NOVELTY LABELS

Our novelty labels are intended to describe the contribution remaining beyond retrieved prior work. We examine this interpretation in two ways: supervised probes predict the stored labels from alternative inputs, and an exploratory blind re-assessment tests responses to controlled changes in evidence.

Prediction from metadata and coverage features. We first measure associations with the continuous novelty score, recomputed from the stored annotation components using Sec. 2.1. Numeric features are target character length, word count, publication date, number of retained priors, mean target–prior semantic similarity, and mean prior coverage score. Categorical features are innovation aspect, contribution type, and task family. We use Spearman ρ for numeric features and Kruskal–Wallis ε^2 for categorical features, with reporting thresholds $|\rho| \geq 0.20$ and $\varepsilon^2 \geq 0.06$. Text length, word count, and publication date have weak associations ($\rho = +0.049$, $+0.058$, and -0.130); all three categorical effects are $\varepsilon^2 \leq 0.033$. Prior count, mean semantic similarity, and mean coverage have negative associations ($\rho = -0.286$, -0.232, and -0.471).

To test their combined predictive value, we fit two class-balanced logistic regression classifiers to the stored four-class novelty labels. The first uses all features listed above except mean coverage; the second adds mean coverage to that same feature set.

Quadratic weighted kappa (QWK) against the stored labels rises from 0.227 without coverage to 0.539 with coverage. This comparison shows the additional predictive value of the coverage annotation beyond the other recorded features.

Prediction with retrieved and random priors. We next train a text probe to test whether the choice of prior content affects prediction of the same stored labels. In the TRUE condition, the input concatenates the target text and up to five retained prior texts. In the RANDOM condition, it concatenates the same target with up to five priors sampled uniformly without replacement from the deduplicated prior pool, restricted to dates earlier than the target.

The probe is TF–IDF followed by class-balanced logistic regression. TF–IDF uses lowercased unigrams and bigrams, sublinear term frequency, a minimum document frequency of two, and at most $200{,}000$ features.

Test QWK is 0.196 with retrieved priors and 0.005 ± 0.021 with random priors (mean and standard deviation across seeds), a gap of 0.192. Retrieved prior content therefore carries more information about the stored labels than random earlier content under this probe.

Response to evidence covering the target contribution. The prediction experiments keep the labels fixed. We now test whether novelty judgments respond appropriately when the evidence changes. If an earlier study already contains the target contribution, the assigned novelty should decrease. We sample 30 targets across the four grades and construct one synthetic prior for each target by copying its technical text and assigning a date one month earlier. After adding it to the evidence set, we use `deepseek-flash` to perform a blind re-assessment of pair coverage, joint coverage, residual contribution, evidence sufficiency, and surface-level difference. The judge sees only the texts, dates, and anonymized prior identifiers; the novelty score and grade are computed using the rule of Sec. 2.1. Synthetic priors have dedicated identifiers and are marked `synthetic_control` in the experiment records. They do not refer to real papers, authors, or DOIs and are excluded from the released dataset.

Each modified target is compared with the medians of five independent blind assessments of its original evidence set to account for variation across repeated assessments. We evaluate recognition of the added prior as large coverage, changes in joint coverage and residual contribution, and decreases in the novelty score and grade. Ties are excluded from the exact two-sided sign tests.

The added prior is labeled large coverage in all 29 comparable cases. Joint coverage increases and residual contribution decreases in 27 cases each (sign test $p = 2.2 \times 10^{-7}$ for each component). The novelty score decreases in $27/29$ cases ($p = 1.6 \times 10^{-6}$), and the novelty grade decreases for 82.8% of targets. These results show that the blind judge lowers novelty when the supplied evidence covers the target contribution.

Stability under changes that preserve the contribution. A novelty judgment should also remain stable when the evidence changes without altering the substantive contribution of the target. For the same 30 targets, we reorder the priors, duplicate existing priors, or add unrelated papers. The novelty grades before and after these changes show agreement measured by quadratic weighted Cohen's κ (Cohen, 1960) of 0.710, 0.812, and 0.889, respectively. These results complement the coverage

Table 4: Symbols used in Stages A–C. Indices: k ranges over retained priors, m over semantic units, j over target tokens, and $i < j$ over unit pairs.

Symbol	Meaning	Size
Sizes and index sets		
d	encoder dimension	768
K	retained priors per target	5
M	semantic units	8
N	target tokens	—
n	retained priors, $\lvert\mathcal{E}_x\rvert$	$\leq K$
$\mathcal{E}_x$	evidence set of target x	—
$\mathcal{V}$	non-padding target positions	—
Stage A: historical evidence		
$\boldsymbol{\pi}_k^{\text{pair}}$	pair coverage distribution	$\mathbb{R}^4$
s_k	pair coverage score	$[0,1]$
$\boldsymbol{f}_k$	pair coverage features	$\mathbb{R}^{11}$
$\boldsymbol{h}_k^J$	pair representation (joint model)	$\mathbb{R}^d$
$\boldsymbol{z}_J$	bag representation	$\mathbb{R}^d$
$\boldsymbol{\pi}^J$	joint coverage distribution	$\mathbb{R}^4$
C_J	scalar joint coverage	$[0,1]$
$\boldsymbol{z}_G$	global context	$\mathbb{R}^d$
Stage B: residual decomposition		
$\boldsymbol{v}_k, \widetilde{\boldsymbol{v}}_k$	pair representation, augmented	$\mathbb{R}^d$
$\boldsymbol{h}_j$	target token state	$\mathbb{R}^d$
$\boldsymbol{q}_m$	learned unit query	$\mathbb{R}^d$
$\boldsymbol{u}_m$	semantic unit	$\mathbb{R}^d$
ϕ_m	unit activity	$[0,1]$
A_{km}	support of prior k for unit m	$[0,1]$
c_m	unit coverage (noisy-OR)	$[0,1]$
r_m	unit residual weight	$[0,1]$
$\widehat{C}_U$	reconstructed joint coverage	$[0,1]$
$\boldsymbol{t}_{ij}$	interaction representation	$\mathbb{R}^d$
g_{ij}	interaction strength	$[0,1]$
$\bar{A}_{ij}$	co-support, $\max_k A_{ki}A_{kj}$	$[0,1]$
r_{ij}	interaction residual weight	$[0,1]$
η_{ij}	interaction attention	$[0,1]$
$\boldsymbol{z}_U, \boldsymbol{z}_I, \boldsymbol{z}_C$	unit, interaction, covered pooling	$\mathbb{R}^d$
Stage C: novelty prediction		
$\hbar_x$	target representation	$\mathbb{R}^d$
γ	context gate	$[0,1]$
$\boldsymbol{z}_R$	fused residual	$\mathbb{R}^d$
$\boldsymbol{h}$	shared representation	$\mathbb{R}^{4d}$

experiment: the judgments respond to evidence that reduces the remaining contribution and remain comparatively stable under changes that preserve it.

C Method Details

C.1 Notation and dimensions

Table 4 collects the symbols of all three stages. Every vector representation lies in $\mathbb{R}^d$, so the MLPs take inputs of dimension $4d+11$ for f_A, $4d$ for f_I and $3d$ for f_R, and the shared representation $\boldsymbol{h}$ has dimension $4d$. Section 3 defines the symbols themselves; the remainder of this appendix supplies the components each stage leaves unspecified.

C.2 Historical coverage models

Pair coverage and evidence selection. For each target–prior pair (x, p_k), the pair coverage model estimates how much of the target the prior accounts for. A cross-encoder predicts relatedness a_k,

coverage conditional on relatedness b_k, and large coverage conditional on coverage d_k, which factorize into a distribution over four ordered classes, namely unrelated, related but uncovered, partially covered, and largely covered:

$$\boldsymbol{\pi}_k^{\text{pair}} = [1 - a_k,\ a_k(1 - b_k),\ a_k b_k(1 - d_k),\ a_k b_k d_k]. \tag{1}$$

Indexing the classes from 0 to 3, the model scores the pair as $s_k = \frac{1}{2}\pi_{k,2}^{\text{pair}} + \pi_{k,3}^{\text{pair}}$, counting partial coverage at half the weight of large coverage. The pipeline ranks temporally admissible candidates by $s_k + 0.02\ell_k/3$, where ℓ_k is the predicted ordinal coverage label, and retains at most K of them. The retained priors and their pair coverage features form the evidence inputs to the subsequent stages.

Pair feature vector. The features carried into Stages B and C are

$$\boldsymbol{f}_k = \left[\, s_k;\ s_k + 0.02\ell_k/3;\ \boldsymbol{\pi}_k^{\text{pair}};\ a_k;\ a_k b_k;\ d_k;\ r_k/K;\ \mathbb{1}[\text{prior text available}] \,\right] \in \mathbb{R}^{11}, \tag{2}$$

where r_k is the rank of the prior within $\mathcal{E}_x$. Through Eq. (1), the first seven entries are deterministic functions of (a_k, b_k, d_k), so $\boldsymbol{f}_k$ carries three degrees of freedom from the pair model alongside the two ranking quantities and the availability flag. We keep the redundant parameterization because the downstream projections consume these quantities at different scales.

Joint coverage. Different priors may cover different parts of the target, so a single pair score understates what the evidence establishes together. We therefore estimate joint coverage with a multiple-instance model that treats each pair (x, p_k) as an instance and the retained evidence set as a bag (Ilse et al., 2018). Let $n = |\mathcal{E}_x|$ denote the number of retained priors. A shared cross-encoder maps each pair to a representation $\boldsymbol{h}_k^J$, from which the model computes attention weights β_k and internal evidence scores e_k:

$$\beta_k = \text{softmax}_k(\boldsymbol{w}_a^\top \boldsymbol{h}_k^J + b_a), \qquad e_k = \sigma(\boldsymbol{w}_e^\top \boldsymbol{h}_k^J + b_e). \tag{3}$$

For a nonempty evidence set, noisy-OR pools the evidence scores as $e_\cup = 1 - \prod_{k=1}^{n}(1 - e_k)$. Four statistics summarize the scores: that aggregate, their maximum and mean, and the fraction of the K evidence slots that are filled. We concatenate these statistics with the attention-pooled and elementwise maximum-pooled pair representations:

$$\boldsymbol{v}_E = \left[e_\cup,\ \max_k e_k,\ \frac{1}{n}\sum_{k=1}^{n} e_k,\ \frac{n}{K} \right], \tag{4}$$

$$\boldsymbol{z}_J = f_J\left(\left[\sum_{k=1}^{n} \beta_k \boldsymbol{h}_k^J ; \max_k \boldsymbol{h}_k^J ; \boldsymbol{v}_E \right] \right), \tag{5}$$

where f_J is an MLP. Attention pooling aggregates evidence across priors, while maximum pooling retains the strongest feature values; the noisy-OR aggregate enters only as one of the four statistics, and the prediction itself is learned from the full bag representation. A hierarchical head maps $\boldsymbol{z}_J$ to three probabilities: any coverage, substantial coverage conditional on any coverage, and mostly covered conditional on substantial coverage. Under the same factorization as Eq. (1), these probabilities yield $\boldsymbol{\pi}^J$ over four joint-coverage classes mirroring the pair classes: *not*, *weakly*, *partially*, and *mostly* covered. The scalar estimate is $C_J = \sum_{\ell=0}^{3} \mu_\ell \pi_\ell^J$, where μ_ℓ is the mean annotated coverage score for class ℓ on the training split. Both C_J and $\boldsymbol{\pi}^J$ are supplied as fixed inputs to the residual and novelty modules.

Global context features. The twenty-one scalars projected into $\boldsymbol{z}_G$ are the scaled target timestamp, ten pair-score statistics, and ten joint-coverage features. The pair-score statistics over $\mathcal{E}_x$ are the fraction of the K evidence slots that are filled; the maximum, mean, standard deviation and sum of the s_k; a union score over them; the fractions of retained priors judged covering and judged related; the margin between the top two ranking scores; and the entropy of the pair scores. The joint-coverage features are two scalar coverage estimates, the raw and normalized entropies of $\boldsymbol{\pi}^J$, an indicator for whether a joint prediction is available, the gap between the joint estimate and the pair-score union, and $\boldsymbol{\pi}^J$ itself.

C.3 Residual representations and pooling

This section specifies the encoders, pooling rules, and masking behavior that Section 3.2 leaves implicit.

Pair encoding. The residual cross-encoder shares parameters across all retained priors. For each target–prior pair (x, p_k) it produces token states $\boldsymbol{H}_k \in \mathbb{R}^{L\times d}$, where L is the input sequence length. Masked mean pooling over these states gives the pair representation, which we augment with projected pair coverage features:

$$\boldsymbol{v}_k = \mathrm{MeanPool}(\boldsymbol{H}_k), \qquad \widetilde{\boldsymbol{v}}_k = \boldsymbol{v}_k + f_P(\boldsymbol{f}_k), \tag{6}$$

where f_P consists of a linear map, layer normalization, GELU, and dropout. The target token bank averages target-token states at matching positions across valid pairs, and masked mean pooling over the bank yields the target representation $\hbar_x$ that Stage C fuses.

Coverage reconstruction. To supervise the unit coverage estimates, we aggregate them into a scalar joint-coverage prediction. The pooling weights depend on both the semantic representation and the activity of each unit:

$$\widehat{C}_U = \sum_{m=1}^{M} \omega_m c_m, \qquad \omega_m = \frac{\phi_m \exp(\boldsymbol{w}_c^\top \boldsymbol{u}_m + b_c)}{\sum_{r=1}^{M} \phi_r \exp(\boldsymbol{w}_c^\top \boldsymbol{u}_r + b_c)}. \tag{7}$$

Annotated coverage and the fixed Stage A prediction supervise the resulting estimate $\widehat{C}_U$. This supervision is aggregate: it constrains the unit coverages without determining the latent support matrix uniquely. Noisy-OR likewise serves as a differentiable rule for aggregating support, and using it does not establish that support from different priors is statistically independent.

Interaction encoding and attention. For each unit pair $i < j$, an MLP f_I encodes the two units and their elementwise comparisons into an interaction representation. A sigmoid head then predicts the interaction strength:

$$\boldsymbol{t}_{ij} = f_I([\boldsymbol{u}_i; \boldsymbol{u}_j; \boldsymbol{u}_i \odot \boldsymbol{u}_j; |\boldsymbol{u}_i - \boldsymbol{u}_j|]), \qquad g_{ij} = \sigma(\boldsymbol{w}_g^\top \boldsymbol{t}_{ij} + b_g). \tag{8}$$

Using the interaction residual weights r_{ij} defined in Stage B, we compute attention over all $\binom{M}{2}$ unit pairs:

$$\eta_{ij} = \mathrm{softmax}_{i<j}\left(\boldsymbol{w}_I^\top \boldsymbol{t}_{ij} + b_I + \log(r_{ij} + \varepsilon)\right), \qquad \boldsymbol{z}_I = \sum_{i<j} \eta_{ij} r_{ij} \boldsymbol{t}_{ij}. \tag{9}$$

The logarithmic bias directs attention toward pairs with larger residual weights. Multiplying by r_{ij} after normalization then restores their scale, so that weak residuals contribute less than the softmax alone would allow.

Masking and edge cases. Padded evidence slots carry zero support: the evidence-availability flag masks the support matrix, which is clamped to $[0, 0.995]$ before use. Both aggregations remain well defined when fewer than K priors are retained. The noisy-OR product runs over the retained priors only, while the co-support maximum is taken over all K slots, so $\bar{A}_{ij} = 0$ when none is retained.

C.4 Residual fusion, prediction heads, and decision rules

This section specifies the gate, the heads, and the decision rules that Section 3.3 states only in outline.

Context gate. The fusion gate conditions the contribution of global context on the two residual representations and the uncertainty of the joint-coverage prediction. Let H_J denote the normalized entropy of $\boldsymbol{\pi}^J$. The gate is

$$\gamma = \sigma\left(f_\gamma([\boldsymbol{z}_U; \boldsymbol{z}_I; \boldsymbol{z}_G; H_J])\right). \tag{10}$$

Here f_γ is a two-layer MLP with a $d/2$ bottleneck and a scalar output, so γ scales $\boldsymbol{z}_G$ uniformly rather than gating it per dimension. The gated context $\gamma \boldsymbol{z}_G$ enters residual fusion alongside $\boldsymbol{z}_U$ and $\boldsymbol{z}_I$, while $\boldsymbol{z}_G$ reaches the shared representation on an ungated path of its own and thereby also reaches the prediction heads.

Auxiliary predictions. The auxiliary heads use the shared representation $\boldsymbol{h}$ to predict substantive difference $\widehat{d}$, surface change $\widehat{s}$, evidence sufficiency $\widehat{e}$, and residual-type probabilities $\boldsymbol{\pi}^T$. The first three outputs are sigmoid scores; the type prediction uses the hierarchy described below.

Residual types. The type hierarchy distinguishes a minor variant from six substantive families: conceptual innovation, mechanism or system, task or application, training or evaluation, data-related innovation, and combination. Conditional fine heads yield eleven types in total. The minor/substantive gate additionally reads $\widehat{d}$, $\widehat{s}$, and $\max_{i<j} r_{ij}$. The probability assigned to the combination type is multiplied by $0.70 + 0.30\sigma(f_{\text{comb}}(\boldsymbol{z}_I))$ before the hierarchical distribution is renormalized. The final type distribution is a convex mixture of the hierarchical and flat predictions, with flat weight 0.15.

Novelty branches. All novelty branches receive the prediction features

$$\boldsymbol{\xi} = [\boldsymbol{h}; \widehat{d}; \widehat{s}; \widehat{e}; \boldsymbol{\pi}^T; C_J; \boldsymbol{\pi}^J]. \tag{11}$$

We encode LOW, WEAK, MODERATE, and HIGH as 0, 1, 2, and 3, respectively. A direct linear classifier and a two-layer MLP classifier map $\boldsymbol{\xi}$ to class distributions $\boldsymbol{p}^D$ and $\boldsymbol{p}^R$.

The ordinal branch models cumulative probabilities of reaching each grade threshold. It produces three sigmoid gates g_1, g_2, g_3 from $\boldsymbol{\xi}$ and defines

$$q_1 = g_1, \quad q_2 = g_1 g_2, \quad q_3 = g_1 g_2 g_3, \qquad \boldsymbol{p}^O = [1 - q_1,\ q_1 - q_2,\ q_2 - q_3,\ q_3]. \tag{12}$$

Here q_ℓ estimates the probability that $y \geq \ell$, with $q_1 \geq q_2 \geq q_3$ by construction.

Probability fusion. We aggregate the three branch distributions through a convex mixture:

$$\boldsymbol{p} = \lambda_D \boldsymbol{p}^D + \lambda_O \boldsymbol{p}^O + \lambda_R \boldsymbol{p}^R, \qquad \lambda_D + \lambda_O + \lambda_R = 1, \quad \lambda_D, \lambda_O, \lambda_R \geq 0. \tag{13}$$

The fused distribution provides the initial grade for the refinement below.

Moderate-to-high refinement. A separate binary head $h_{MH} = \sigma(f_{MH}(\boldsymbol{\xi}))$ scores the distinction between moderate and high novelty. At inference, let $y_0 = \arg\max_\ell p_\ell$ for the mixture in Eq. (13), and define

$$h_{\text{high}} = \alpha_D p_3^D + \alpha_O p_3^O + \alpha_{MH} h_{MH}, \qquad \widehat{y} = \begin{cases} 3, & y_0 = 2 \text{ and } h_{\text{high}} \geq \theta_{MH}, \\ y_0, & \text{otherwise.} \end{cases} \tag{14}$$

where $\alpha_D, \alpha_O, \alpha_{MH} \geq 0$ and $\alpha_D + \alpha_O + \alpha_{MH} = 1$. Thresholding the weighted score, rather than requiring separate agreement from all three heads, promotes a MODERATE prediction to HIGH whenever the heads are collectively confident, even when each of them favors a different grade. We select the mixture weights and the refinement threshold on development data and fix them for testing; Appendix C.5 reports the values.

Continuous outputs. A separate sigmoid head predicts a continuous novelty score. Probability fusion and the moderate-to-high refinement alone determine the grade at inference, so this score never enters the decision rule.

C.5 Training and implementation details

Optimization scope. We initialize the residual encoder from the Stage A joint-coverage checkpoint and jointly train it with the semantic decomposition, residual fusion, and prediction modules. The Stage A pair and joint coverage models remain fixed.

Objective. Let $\mathcal{A} = \{\Delta, S, E, C, \nu\}$ index the annotated scalars: the substantive difference, surface change, and evidence sufficiency read off the auxiliary heads, the reconstructed coverage $\widehat{C}_U$, and the continuous novelty score. Their losses combine scalar regression with auxiliary ranking and coverage consistency terms, as specified below. The full objective is

$$\mathcal{L} = \sum_{a \in \mathcal{A}} \lambda_a \mathcal{L}_a + \lambda_{\text{rank}} \mathcal{L}_{\text{rank}} + \lambda_{\text{nov}} \mathcal{L}_{\text{nov}} + \lambda_{\text{type}} \mathcal{L}_{\text{type}} + \lambda_{\text{dec}} \mathcal{L}_{\text{dec}} + \lambda_{\text{var}} \mathcal{L}_{\text{var}}, \tag{15}$$

Here $\mathcal{L}_{\text{nov}}$ supervises the three novelty branches, their fused distribution, and the moderate-to-high head; $\mathcal{L}_{\text{type}}$ supervises residual categories and nontrivial combinations; and $\mathcal{L}_{\text{dec}}$ and $\mathcal{L}_{\text{var}}$ regularize the decomposition. Each coefficient λ weights its term in the sum.

Novelty and residual-type supervision. $\mathcal{L}_{\text{nov}}$ trains the direct and MLP classifiers with class-weighted cross-entropy, the ordinal outputs against $\mathbf{1}[y \geq \ell]$, $\ell = 1, 2, 3$, with weighted binary cross-entropy, and the fused distribution with a negative log-likelihood. The moderate-to-high head sees only moderate and high examples and adds a positive–negative margin to its binary cross-entropy, weighting moderate examples scored as high more heavily. $\mathcal{L}_{\text{type}}$ combines the hierarchy's conditional losses, a flat classification loss, binary supervision for nontrivial combinations, and consistency penalties tying the combination branch to the type predictions. The novelty score regresses ν^*, the continuous target that $\text{raw}(x)$ in the novelty rule of Sec. 2.1 forms from the annotated coverage, substantive difference, evidence sufficiency and surface change of the target, and that the dataset thresholds into the four grades.

Scalar regression and ranking. Let ℓ_δ denote smooth-L_1 loss and w_x the example weight. For $a \in \mathcal{A}$, the regression component is $\mathcal{L}_a^{\text{reg}} = \sum_x w_x \ell_\delta(\widehat{a}_x - a_x^*) / \sum_x w_x$, where a_x^* is the annotation and $\widehat{a}_x$ the prediction. The ranking term takes within-minibatch pairs separated by $|a_x^* - a_{x'}^*| \geq 0.05$ and contributes $[0.02 - \text{sgn}(a_x^* - a_{x'}^*)(\widehat{a}_x - \widehat{a}_{x'})]_+$, weighted by $\sqrt{w_x w_{x'}}$; it is dropped when a minibatch holds no eligible pair.

Regression supervises the numerical values, while ranking explicitly encourages separation in the annotated order. The fixed margin also sets a minimum desired separation between predictions. Ranking terms for Δ, S, E, and C are collected in $\mathcal{L}_{\text{rank}}$; the continuous novelty loss $\mathcal{L}_\nu$ combines its own regression and ranking terms. The continuous score provides auxiliary training supervision and does not determine the inference-time novelty grade.

Coverage consistency. Beyond its level loss against the annotated joint value C^*, $\mathcal{L}_C$ adds a consistency loss against the upstream score C_J at relative weight 0.2, reweighted by $1 - H_J$ so that examples with uncertain upstream coverage count for less. This constrains reconstructed joint coverage; no unit-level support matrix is supplied as a target.

Decomposition regularization. $\mathcal{L}_{\text{dec}}$ combines three terms: attention regularization penalizes the squared entries of $\widetilde{\alpha}\widetilde{\alpha}^\top - I$ for the row-normalized attention matrix $\widetilde{\alpha}$, plus half the mean off-diagonal cosine similarity; query diversity penalizes the mean squared cosine similarity between distinct queries; and a usage penalty $D_{\text{KL}}(\boldsymbol{\rho} \| \boldsymbol{u})$, with $\rho_m = \sum_j B_{mj} / \sum_{r,j} B_{rj}$ and $\boldsymbol{u}$ uniform over the M units, keeps units from going unused. The variance floor carries its own coefficient,

$$\mathcal{L}_{\text{var}} = \mathbb{E}_x \big[[\nu - \text{Var}_m(c_{x,m})]_+ / \nu \big], \tag{16}$$

where ν here denotes the positive variance floor. This term acts on unit coverages rather than token assignments and discourages identical coverage estimates within an example. It imposes a preference for differentiated unit coverages, rather than providing additional coverage labels.

Curriculum and feature perturbation. For a randomly sampled fraction of training examples, values derived from annotated joint coverage replace the joint-coverage features and the associated probability and entropy features with probability 0.50 in epoch 1, 0.20 in epoch 2, and 0 thereafter. Development and test runs always use the upstream joint predictions. Selected joint-coverage features on the global-context path also receive Gaussian noise of standard deviation 0.03 and independent feature dropout at probability 0.10, which leaves untouched both the direct coverage inputs to the novelty heads and the entropy input to the fusion gate.

Input handling. The input adapter reconstructs pair probabilities missing from a ranking record from the available coverage label and score; pair labels and scores also appear in the pair text input. Padded priors have zero support. When the evidence set is empty, the input contains the target text and all prior support is masked out. Table 5 summarizes the model and training settings.

Table 5: Model and training settings for residual decomposition and novelty prediction.

Setting	Value
Encoder	DeBERTa-v3-base
Hidden dimension d	768
Maximum pair-input length	384 tokens
Maximum retained priors K	5
Semantic units M	8
Dimension of each facet embedding	32
Support upper bound	0.995
Annotated-feature replacement probability	0.50, 0.20, then 0
Context-feature noise standard deviation	0.03
Context-feature dropout probability	0.10
Training epochs / selected epoch	8 / 5

D Experimental Details

D.1 Setup

Splits. The split is a date-ordered $13{,}815/3{,}944/3{,}352$ partition of the $21{,}111$ idea-level samples in IdeaDelta.

Configuration. The default EGRD configuration uses a `deberta-v3-base` backbone (He et al., 2023), 8 epochs, batch size 2, learning rate 1.5×10^{-5} with an encoder learning-rate ratio of 0.2, eight semantic units with slot competition enabled, and a fixed joint-prediction checkpoint supplying the encoder initialisation. The checkpoint reported in the main text is selected on dev; test is traversed once, at final evaluation, and is not used to select a checkpoint.

Dense retrieval baseline. The dense-RAG rows of Table 1 replace the stored candidate lists with retrieval performed at query time (Karpukhin et al., 2020). We embed every canonical idea with `bge-large-en-v1.5` (Xiao et al., 2024), index the ideas whose date precedes the target's, query the index with the target text, and pass the top five hits to the same single-call prompt the *direct* rows use. Nothing else changes: the same backbones, the same prompt, the same labels and the same scoring script. Each domain is indexed separately, so a target retrieves within its own domain as every other row does. Inside that index the date is the only constraint applied: the search ranges over all aspects, whereas the stored candidate lists come from aspect-restricted nearest-neighbour recall together with the cross-type pass of §2.

Metrics. We use four evaluation metrics, computed over the four ordered grades LOW, WEAK, MODERATE, HIGH, indexed 0 to 3. **Accuracy** is the fraction of ideas whose predicted grade equals the gold grade. It weights every example equally and does not distinguish errors by grade distance, so frequent classes can dominate it. **Macro-F1** averages the per-class F_1 over the four grades with equal weight, which charges a system for abandoning a small class rather than letting the large ones absorb the loss. **QWK** is quadratic weighted kappa (Cohen, 1968),

$$\kappa_w = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}}, \qquad w_{ij} = (i-j)^2/9,$$

where O is the confusion matrix and E the matrix expected from the two marginals if prediction and gold were independent. QWK measures chance-corrected ordinal agreement, penalizing distant grade errors more heavily than near misses. With a nonzero denominator, it equals 1 for exact agreement, 0 when observed weighted disagreement matches the chance reference, and is negative when disagreement exceeds that reference. A constant predictor has QWK 0 on our evaluation sets; independent random predictions need not yield exactly 0 in a finite sample. QWK depends on the assigned grades and their marginal distributions, so it does not isolate ranking quality from grade calibration. **MAE** (mean absolute error) measures the average absolute difference between predicted and gold novelty grades:

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\widehat{y}_i - y_i|,$$

where $\widehat{y}_i$ and y_i are the predicted and gold grade indices in $\{0, 1, 2, 3\}$, respectively, and N is the number of evaluated ideas. Lower values indicate smaller grade errors. Among constant predictions, a median of the gold-label distribution minimizes MAE; this does not characterize the best input-dependent predictor. We therefore interpret MAE alongside the other metrics. Every table in the paper reports these four. On RINoBench the expert 1–5 scale is folded to the same four levels before any of them is computed, so every number in Table 2 is on the scale described here.

Per-domain metrics. Per-domain figures are recomputed from the stored predictions rather than read from the trainer, and reconcile with the trainer's own pooled figures to within 5×10^{-5} on all four metrics.

D.2 RINoBench transfer protocol

Data and labels. RINoBench (Schopf & Färber, 2026) draws research ideas from OpenReview submissions across machine learning and pairs each with admissible prior work. Its 277 test ideas are rated by domain experts for novelty on a 1–5 scale, providing human judgments beyond the model-generated labels and three-domain coverage of IdeaDelta. We merge ratings 4 and 5 to obtain four ordered grades, placing 41.5% of the test ideas in the top class.

Evidence and inference. We use the checkpoint reported in Table 1 without retraining or fine-tuning on RINoBench. Each target has 25 benchmark-supplied prior ideas, compared with five in the IdeaDelta protocol. To retain the trained input budget of $K{=}5$, we split the 25 priors into five blocks of five and run the network once per block. Each pass produces a continuous substantive-difference score; the five scores are aggregated into one, which is mapped to four ordered grades using three thresholds.

Baselines and scoring. We rerun Qwen3.5, GraphMind, and the UKPLab three-stage pipeline on the same benchmark evidence and include the predictions released by the RINoBench authors. DeepReviewer-7B is evaluated using its local weights. All systems in Table 2 are scored on the same four-grade scale using the evaluation script that produces Table 1.

D.3 Parameter sensitivity

Each sweep varies one parameter from the default configuration of Table 1 and holds the rest fixed: the upstream pair and joint products, the encoder initialisation, the evaluation files and the seed. Every point is a single run at seed 173, except for the seed sweep below.

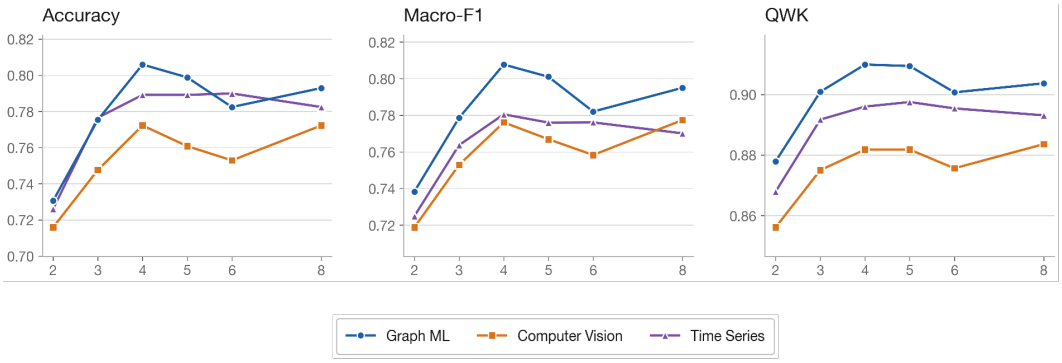

Figure 6: Sensitivity to the evidence budget K in the three domains. The default is $K{=}5$, the checkpoint of Table 1. K specifies the number of input slots, including padding when fewer priors are available.

Evidence budget: gains flatten beyond four input slots. Figure 6 varies the input-slot budget K over $K = 2, 3, 4, 5, 6, 8$. Accuracy rises sharply as the budget increases from two to four slots, then levels off across all three domains. This shared pattern shows that an overly small evidence budget restricts the model's ability to assess a target against prior work. Expanding the budget initially provides a more informative basis for comparison, whereas further expansion produces only small changes in prediction quality under this setup. This plateau mainly reflects evidence availability:

retrieval yields around five prior ideas per target on average, so most additional slots beyond this budget are padded rather than filled with new evidence. Increasing K therefore adds little actual evidence for most targets. The default $K=5$ lies in this stable region, supporting its use as a compact evidence budget that captures most of the observed performance gains.

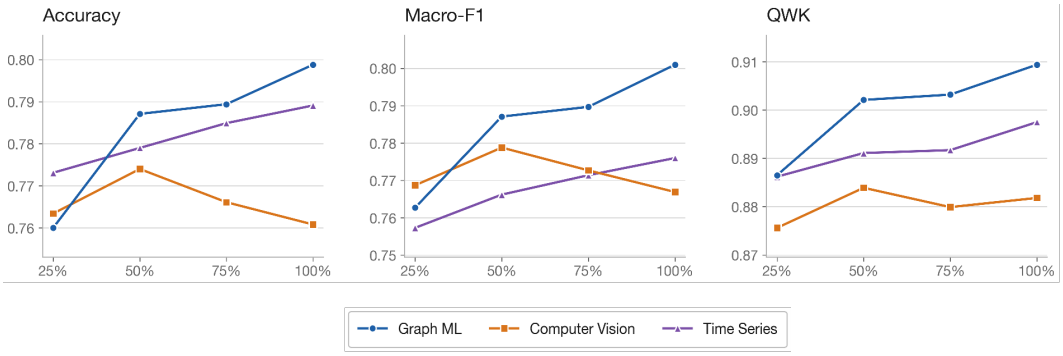

Figure 7: Sensitivity to the amount of EGRD training data in the three domains. Development and test sets stay fixed, and 100% is the default.

Training-set size: additional data generally improves performance. Figure 7 varies the number of training ideas over 25%, 50%, 75% and 100% of the training split, drawn as nested stratified prefixes so that consecutive points add samples without reordering what came before. Overall, increasing the training data generally improves performance, suggesting that exposure to more target–prior relationships helps the model learn comparison patterns and refine its novelty judgments. The gains remain modest across the tested fractions, indicating that smaller training subsets already provide a useful basis for learning, with additional data bringing further refinement.

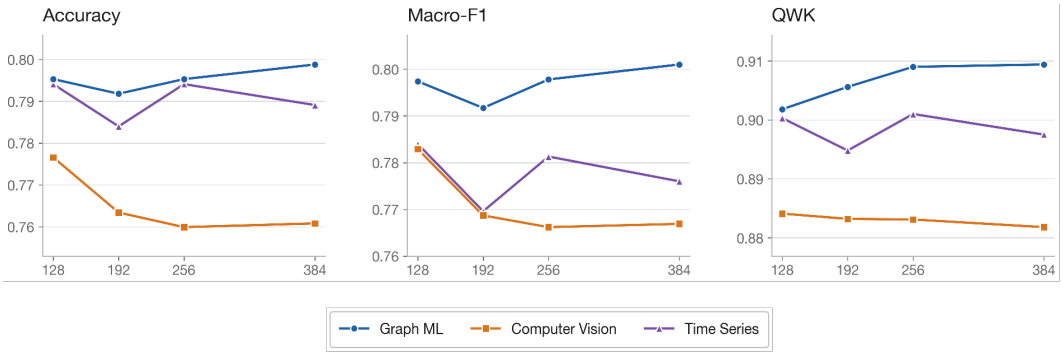

Figure 8: Sensitivity to encoder input length on the three domain test splits. The default is 384, and every setting scores the same test ideas and labels.

Encoder input length: limited sensitivity over the tested range. Figure 8 varies the encoder input length over 128, 192, 256 and 384 tokens. Performance remains broadly stable across these settings, with longer inputs providing no consistent gain. This pattern is consistent with the concise contribution statements used to represent ideas: the information needed for target–prior comparison is concentrated in a compact description. Shorter inputs can therefore provide an effective basis for identifying semantic overlap and substantive differences, while extending the input budget offers limited additional benefit. This stability suggests that the model can operate effectively with a compact input budget.

Table 6: Per-domain seed variability. Entries are sample standard deviations across seeds 173, 174 and 175 with identical data, code and configuration.

Domain	Accuracy	Macro-F1	QWK	MAE
Graph ML	0.0066	0.0050	0.0048	0.0067
Computer vision	0.0040	0.0021	0.0032	0.0027
Time series	0.0050	0.0045	0.0031	0.0051

Seed stability: the reported model is not a selected seed. Table 6 varies nothing but the random seed, running the default configuration on seeds 173, 174 and 175 with identical data and code and reporting the per-domain sample standard deviations. Performance remains stable across seeds, with fluctuations much smaller than the losses caused by the ablations in Figures 3 and 4. This consistency supports the reproducibility of the reported results and indicates that the contribution of the model components exceeds ordinary training variability. The three parameter sweeps above should therefore be interpreted through their overall trends rather than small differences between adjacent settings.

D.4 Case Study Details

Construction of Figure 5. Panel (a) is read off the forward pass of the trained model and fits nothing. For the other two panels the test ideas are split $80/20$ with the three cases held out, and every supervised step is fit on the 80% part alone. Panel (b) standardises $\boldsymbol{z}_R$ on that part, ranks its coordinates by univariate F-statistic against the grade, and shows the 140 that separate it best. Panel (c) reduces $\boldsymbol{z}_R$ to 50 principal components and projects it onto the first two axes of a shrinkage linear discriminant fit on the four grades. The probe reported in the panel is separate from both: a five-fold cross-validated multinomial logistic regression on standardised $\boldsymbol{z}_R$, scored out of fold.

Table 7 gives the three targets of §4.4 and the five retained priors of each, with the date that makes the prior admissible, the coverage label the annotation carries, and the support A_k the model assigns. Ideas are stored as single contribution statements, so each row is the whole of the text the model reads for that idea.

E Limitations

Two bounds on the claims of this paper are worth stating directly.

Three domains. IdeaDelta is drawn from graph machine learning, computer vision and time series, and the RINoBench transfer (§4.2) indicates what lies outside them: roughly 30 of its 277 test ideas fall in these three fields, and the model reads the remainder without ever having been trained on their literature. Coverage of a field is what the decomposition needs, since a prior can only support a unit of a target the model is equipped to read, so we do not claim the reported Accuracy outside the three.

Transfer to expert judgment is partial. The transfer experiment of §4.2 shows the frozen checkpoint leading on the rank-sensitive metrics at 0.2245 QWK, but 0.2245 is a modest level of agreement, and Accuracy adds nothing to it: our row reaches exactly the constant predictor's 0.4152 with a prediction that is not constant. What transfers is the ordering of ideas, not the grade an idea receives, and the label source, the domain mix and the folded scale all shift at once, so the result is partial agreement with expert judgment rather than evidence that the two scales coincide.

F AI Use Statement

We used generative AI tools to assist with the construction of IdeaDelta, including extracting and filtering research ideas, assessing target–prior coverage, and annotating research novelty, as described in Section 2 and Appendix B. The resulting labels are model-generated, and we manually inspected samples from each stage of the construction pipeline. We also used generative AI tools to polish the language of the manuscript and improve its clarity and readability. The authors take responsibility for the final content of this work, including its text, claims, and AI-assisted artifacts.

G Ethics Statement

This work studies research novelty relative to a retrieved set of prior works. The labels in IdeaDelta are model-generated assessments rather than authoritative judgments of scientific merit. Incomplete retrieval, errors in idea extraction, and biases of the annotation model may affect these

Table 7: The three case-study targets and their retained priors, verbatim. Coverage labels are *large*, *partial*, *related* and *none*; A_k is the support the model assigns to that prior.

Idea	Date	Cover	A_k	Text
Graded low				
CI_26954	2026-08	—	—	Propose an action-interleaved transformer architecture that explicitly models state transitions by interleaving state and action tokens in causal order to learn controllable dynamics.
CI_14929	2024-01	large	0.838	Introduce specialized action tokens to denote embodied interactions and state tokens to represent subsequent multisensory observations, enabling end-to-end instruction tuning for active exploration loops.
CI_17158	2025-02	partial	0.447	Employ a causal transformer with a query-based parallel representation updating mechanism to jointly model the dynamics of RGB-D data, 3D flow, and robotic actions in an end-to-end manner.
CI_13226	2022-03	partial	0.507	Design an action module that learns to predict state transitions in the latent space of the environment representation in an unsupervised manner to enable interactive control.
CI_8741	2023-06	related	0.070	Propose a generative pre-trained Transformer-based dynamics model to learn soft object behavior from large-scale interaction data.
CI_10777	2018-01	none	0.003	Propose a unified framework that integrates a temporal Conditional Random Field with discriminative sparse coding to jointly model action transitions and learn robust frame-level representations for fine-grained segmentation.
Graded moderate				
CI_18669	2026-01	—	—	Introduce token-level history caching to reuse previously generated image tokens directly, preventing identity drift caused by repeated image reconstruction in multi-round interactions.
CI_14692	2024-01	partial	0.523	Propose an adaptive token propagation mechanism that dynamically pauses and restarts tokens across layers to enable reuse of previously discarded information, addressing the inefficiency of permanent token removal.
CI_17874	2025-03	related	0.026	Propose a content-aware token aggregation mechanism that shares token centers across all image tokens and updates them exclusively during the training phase to eliminate inference overhead while capturing long-range dependencies.
CI_15897	2024-03	related	0.043	Propose a nested tokenization framework that hierarchically organizes image patches into region-level and global-level tokens to enable end-to-end processing of extremely large images without quadratic memory growth.
CI_12981	2022-02	related	0.071	Propose a token reorganization mechanism that dynamically preserves attentive image tokens and fuses inattentive ones into a single token during the feed-forward process to reduce computational redundancy.
CI_14768	2024-01	none	0.006	Introduce an iterative refinement strategy using weak noise levels to correct visual artifacts from 3D projection while preventing structural distortion and identity loss.
Graded high				
CI_18985	2026-02	—	—	Design a policy optimization algorithm that utilizes triplet-based trajectory mixing and contrastive response maximization to explicitly address high intra-class variance and low inter-class variance in fine-grained tasks.
CI_3103	2025-11	related	0.005	Introduce a reward aggregation strategy for deep reinforcement learning that averages returns over equivalent expression sequences to reduce training variance.
CI_18341	2026-01	related	0.007	Develop a unified reinforcement fine-tuning framework that integrates multi-reward normalization and length-constrained ensemble rewards to optimize fine-grained visual classification while suppressing excessive textual reasoning.
CI_17358	2025-02	none	0.002	Develop a reward-driven training strategy that uses student performance improvements and reduced teacher-student discrepancies as feedback signals to optimize the weight generation agent via policy gradient.
CI_14068	2023-02	none	0.002	Introduce a center-concentrated multi-prototype generation strategy to mitigate intra-class sparsity by minimizing intra-class distances and maximizing inter-class separation.
CI_12524	2021-03	none	0.003	Propose a quality estimation mechanism that calculates the distribution distance between intra-class and inter-class similarities to generate pseudo-labels for unsupervised training.

assessments, and uneven coverage across research fields may lead to unequal reliability. The resulting scores should therefore support human examination of the underlying evidence and should not serve as the sole basis for publication decisions, evaluations of researchers, or allegations of plagiarism. The synthetic priors used in the controlled validation experiment are explicitly marked, do not represent real publications, and are excluded from the released dataset, as described in Appendix B.2.

H REPRODUCIBILITY STATEMENT

Model and training code for all three stages is available for review in an anonymized repository at `https://anonymous.4open.science/r/EGRD_model-86D7`, including the pair-coverage and joint-coverage models of Stage A, the residual decomposition of Stage B, the novelty prediction heads of Stage C, and a single entry point that reproduces the reported training pipeline under the hyperparameters used in our experiments. The repository contains model and training code only; it carries no data, checkpoints, or evaluation outputs. IDEADELTA is released separately, also under an anonymized account for review, at `https://huggingface.co/datasets/holaholaIDEADELTA/IdeaDelta`.