

Token Crit at the FinEval 2026: Role-Specialized, Validation-Gated LLM Workflows for JF-ICR and Multilingual Financial Q&A

Anke Xu

Independent Researcher
ankexu329@gmail.com

Wenbo Cao

Independent Researcher
evancao1429@gmail.com

Abstract

Multi-stage LLM workflows can degrade correct outputs when intermediate analyses revise shared state without constraint. We present validation-gated systems for two FinEval 2026 tasks: Japanese Financial Implicit Communication Recognition (JF-ICR) and multilingual financial Q&A (PolyFiQA). Specialized roles produce structured proposals, while task-specific validators admit only scoped label transitions for JF-ICR or source-linked evidence for PolyFiQA. On the 253-example JF-ICR public development set, the final workflow reaches 62.1% accuracy, compared with 51.0% for the stronger direct LLM baseline and 54.9% for a majority-class baseline. Applying all final-stage proposals would yield a net loss of two correct predictions, whereas gated admission yields a net gain of 11. On 76 PolyFiQA training examples, the final workflow yields a modest overall ROUGE-1 improvement (0.3399 to 0.3425) but raises evidence-section ROUGE-1 F1 from 0.1570 to 0.5132. Together, these results suggest that validation-gated updating is most effective for narrow, checkable changes; open-ended synthesis still requires recall-aware admission and closer alignment between validation and answer generation.

1 Introduction

Complex knowledge work, including financial analysis, compliance review, and risk investigation, often combines specialized steps whose intermediate errors can propagate downstream. The two FinEval tasks we study expose different forms of uncertainty. Japanese Financial Implicit Communication Recognition (JF-ICR) requires distinguishing the direction and strength of indirect corporate responses, whereas PolyFiQA requires coordinating multilingual evidence across news and financial statements (Peng et al., 2026b,a). A single end-to-end LLM call may conflate interpretation, evidence selection, and final resolution, allowing an early error to affect the final output.

Adding roles or model calls is not sufficient: during JF-ICR development, unrestricted task-specific re-analysis sometimes overturned correct holistic predictions. This motivates our central question: how should a multi-role workflow separate responsibility for analyzing a problem from authority to update its shared state?

We address this question with two role-specialized workflows organized around a shared validation-gated update principle (Figure 1). Each specialized role proposes a scoped update, and task-specific validators determine whether it modifies the shared state. JF-ICR admits only scoped label transitions and retains the incumbent label when checks fail; PolyFiQA admits candidate evidence only after source- and relation-specific validation.

Our contributions are: (1) a role-specialized, validation-gated formulation of multi-stage LLM inference; (2) task-specific instantiations as constrained label transitions for ordinal pragmatic classification and evidence admission control for multilingual financial QA; and (3) comparisons with direct and intermediate baselines, including analysis of when proposed updates are accepted, rejected, beneficial, or harmful.

2 Tasks and Related Work

JF-ICR. JF-ICR is part of the Ebisu benchmark for Japanese financial language understanding (Peng et al., 2026b). Given a company question and response, it predicts $y \in -2, -1, 0, +1, +2$, ranging from strong refusal to strong commitment. The task requires recovering pragmatic intent from indirect language, hedging, proposition scope, and commitment strength. Its label space combines broad directional ambiguity with difficult adjacent-label boundaries, motivating scoped rather than unrestricted refinement.

PolyFiQA. PolyFiQA is part of MultiFinBen (Peng et al., 2026a). The FinEval task combines

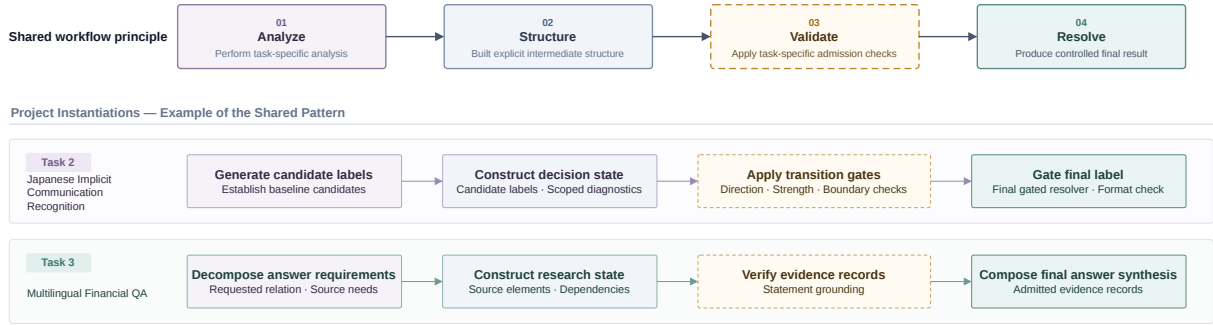


Figure 1: Shared controlled-update abstraction instantiated as gated label transitions for Task 2 and validated, source-linked evidence for Task 3. Rejected updates preserve the incumbent label or unresolved requirement.

SEC financial statements with contemporaneous news in English, Chinese, Japanese, Spanish, and Greek. Evidence may span multiple sources, but topical relevance alone does not establish the relation requested by a question. Numerical support must also match the accounting concept, period, unit, and value.

Related workflows. Prior LLM workflows use staged actions and role specialization (Yao et al., 2023; Chen et al., 2024), while refinement and verification methods revise or check outputs through model or tool feedback (Madaan et al., 2023; Gou et al., 2024; Dhuliawala et al., 2023; Gao et al., 2023). Financial QA benchmarks further emphasize source grounding and numerical operations (Chen et al., 2021; Zhu et al., 2021). Our focus is the downstream authority of role outputs: rather than passing them forward by default, Task 2 admits only specified label transitions, whereas Task 3 admits only validated evidence records.

3 Task-Specific Workflows

3.1 Shared Controlled-Update Principle

Both systems follow the Analyze–Structure–Validate–Resolve pattern shown in Figure 1. *Analyze* delegates a narrow task-specific question to an LLM; *Structure* converts the result into an explicit decision or evidence state; *Validate* determines whether a proposed update is supported and admissible; and *Resolve* commits validated changes while otherwise preserving the incumbent state. The underlying roles, states, and validation rules remain task specific. Task 2 applies this pattern to label transitions, whereas Task 3 applies it to evidence records and constrained synthesis.

3.2 Task 2: Selective Label Resolution

Analyze and Structure. A rubric prompt and commitment decomposition produce complementary five-label candidates. The former assesses the global question–response relation; the latter separates the requested proposition, company stance, and commitment strength. These candidates initialize a decision state with scoped diagnostics for direction, total withholding, commitment strength, polarity and relation, and adjacent-label boundaries.

Validate. When candidates disagree in direction, two diagnostics independently classify the response as commitment, neutral, or refusal. A transition is admitted only when valid records agree, support an existing candidate, and do not indicate a multi-part response. Additional checks distinguish total from partial withholding and restrict strength and boundary repairs to label-local transitions. Unsupported or out-of-scope proposals leave the incumbent unchanged.

Resolve. A deterministic resolver applies admitted transitions sequentially and emits one label from $-2, -1, 0, +1, +2$. Diagnostics thus propose bounded updates rather than restart five-way classification. Appendix A summarizes the gate scopes, rejection behavior, and transition outcomes.

3.3 Task 3: Evidence-Controlled Synthesis

Analyze and Structure. The planner decomposes each question into answer requirements, assigns them to news, financial statements, or both, and records source dependencies. Candidate findings are stored in a structured evidence state, separated from contextual and unresolved material so that topical relevance does not imply answer support.

Validate. News evidence is checked for company attribution, answer-object relevance, and support for the requested relation. Financial evidence is grounded to the required metric, period, unit, and calculation, with deterministic matching and arithmetic where possible. Multilingual findings are normalized only after validation while preserving their grounded meaning.

Resolve. The synthesizer operates only on admitted evidence and is followed by output-level checks for coverage, attribution, numerical grounding, and unsupported claims. Unresolved defects remain unresolved rather than being filled with unsupported content. Appendix A summarizes the evidence-state record types and validation checks.

4 Experimental Setup

4.1 Models and Inference

Task 2. We use API-based inference through OpenRouter with Llama 3.3 70B (Grattafiori et al., 2024) for the rubric-based candidate and Qwen2.5-72B (Qwen Team, 2024) for commitment decomposition. Model outputs are combined through deterministic transition gates and a label resolver.

Task 3. We likewise use API-based inference through OpenRouter. Llama 3.3 70B handles planning, source routing, semantic evidence extraction, and answer synthesis, while Qwen3-Embedding-8B supports semantic retrieval. Deterministic components handle financial matching and calculations, evidence assembly, grounding checks, and output validation. For both tasks, we retain intermediate outputs and inference configurations for reproducibility.

4.2 Data and Metrics

Task 2. We develop on the 253-example labeled public JF-ICR set and generate predictions for the official participant set. The public set is imbalanced: 139 examples (54.9%) are labeled +1, while only two are labeled −2. Exact five-label accuracy is the official metric. As diagnostics, we also report direction accuracy after collapsing the labels into refusal (−2/−1), neutral (0), and commitment (+1/+2), and ordinal mean absolute error (MAE) over the ordered five-label scale. Because the same public set was used for workflow development, we treat its results as in-corpus development estimates.

Task 3. We conduct ablations on the 76 labeled training examples and evaluate the final workflow on the official public and private leaderboard splits. ROUGE-1 F1 is the primary metric, with precision and recall reported as diagnostics. All three are macro-averaged across examples with stemming disabled; consequently, macro F1 need not equal the harmonic mean of the displayed macro precision and recall.

4.3 Comparisons

Task 2. We compare a majority-class baseline, direct single-model prompting, and increasingly constrained workflow variants. The majority baseline assigns +1 to every example, while direct prompting assigns one of the five labels. Gated composition combines Llama and Qwen candidates using direction and total-withholding diagnostics. Strength repair adds a scoped +1/+2 commitment-boundary check, while the final workflow adds deterministic polarity, relation, and adjacent-boundary repairs. These comparisons examine the use of additional model calls as constrained diagnostics rather than unrestricted five-way classifiers.

Task 3. We compare three settings on the same 76 training examples. Direct prompting sends the raw question and query directly to Llama. Filtered Financial Evidence applies planning, source routing, and extraction, then uses a single Llama call over the resulting filtered evidence packet. The final workflow carries the structured evidence state through constrained synthesis and output validation. All settings use the same generation model and provider settings, and gold answers are used only for post-generation evaluation.

5 Results and Analysis

On the Task 2 JF-ICR public set, the majority-class baseline obtains 54.9% accuracy, compared with 49.8% and 51.0% for direct Llama and Qwen prompting. Gated composition reaches 56.1%, and later diagnostics further raise accuracy to 62.1%, 7.2 points above the majority baseline, while reducing ordinal MAE from 0.506 after gated composition to 0.427. Direction accuracy remains at 66.4% through gated composition and strength repair, then increases to 70.0% in the final workflow. The gains therefore reflect both candidate composition and controlled label revision, concentrated in exact-label rather than directional discrimination.

Task 2: JF-ICR Accuracy				
<i>Public-set development</i>				
System		Accuracy	Dir. Acc.	Ord. MAE
Majority baseline (+1)		54.9	66.4	0.510
Direct prompting (Llama)		49.8	68.0	0.590
Direct prompting (Qwen)		51.0	66.4	0.561
Gated composition		56.1	66.4	0.506
Strength repair		57.7	66.4	0.486
Final workflow		62.1	70.0	0.427
<i>Leaderboard (final workflow only)</i>				
Public		100.0	–	–
Private		97.1	–	–

Task 3: ROUGE-1			
<i>Training-set ablations</i>			
System	F1	Precision	Recall
Direct prompting	0.3399	0.3886	0.3437
Filtered financial evidence	0.3415	0.5202	0.3070
Final workflow	0.3425	0.4600	0.3269
<i>Leaderboard (final workflow only)</i>			
Public	0.2969	0.3755	0.2867
Private	0.3062	0.3971	0.2709

Table 1: Task 2 public-set ablations and official leaderboard results (left), and Task 3 training-set ablations and official leaderboard results (right). Only the final workflows were submitted for public and private evaluation.

The final workflow scores 100.0% and 97.1% on the public and private leaderboards. Because the participant set is balanced and substantially shorter than the skewed development set (median combined length: 81 vs. 312 characters), we report these as separate outcomes rather than directly comparable gains.

For Task 3, direct prompting obtains a training-set ROUGE-1 F1 of 0.3399. Filtered Financial Evidence increases precision from 0.3886 to 0.5202, but reduces recall from 0.3437 to 0.3070, yielding an F1 of 0.3415. The complete workflow achieves the highest F1 of 0.3425, recovering part of the lost recall (0.3269) while retaining higher precision than direct prompting (0.4600 vs. 0.3886). This indicates that filtering improves selectivity, while the full workflow partially restores useful coverage.

On the official news-first leaderboard, where financial statements serve as supporting evidence, the final workflow achieves public and private F1 scores of 0.2969 and 0.3062, respectively.

Controlled-update analysis. For Task 2, the final boundary audit proposed 75 label transitions, of which 19 were admitted. Applying all proposals would have reduced the number of correct predictions by two, whereas gated admission yielded a net gain of 11. Because both policies act on the same 75 proposals, this comparison isolates the effect of admission control from the inference required to generate them. The 17 rejected corrections nonetheless illustrate the cost of conservative gating (Table 3).

For Task 3, direct prompting achieves the highest answer F1 (0.2996), while the structured vari-

ants raise evidence F1 (Table 4) substantially from 0.1570 to about 0.513. This pattern suggests that controlling the intermediate evidence state primarily improves evidence selection and grounding rather than answer generation.

6 Conclusion

We presented validation-gated LLM workflows for JF-ICR and PolyFiQA under a shared Analyze–Structure–Validate–Resolve principle. The two tasks clarify where this principle is most effective. For JF-ICR, restricting diagnostics to narrow label transitions raises public-set accuracy from 56.1% after gated composition, close to the 54.9% majority-class level, to 62.1% in the final workflow. Gating the final boundary proposals further changes their effect from a net loss of two correct predictions to a net gain of 11, showing that bounded update authority can protect final decisions when proposed effects are discrete and checkable. For PolyFiQA, the workflow substantially increases evidence-section overlap but provides only a modest overall ROUGE-1 improvement, indicating that a better intermediate evidence state does not automatically produce a better final answer. The shared principle therefore provides an inspectable control abstraction rather than a universal performance guarantee. Open-ended synthesis may require more recall-aware admission, recovery paths for rejected evidence, and closer alignment between validation and answer generation.

Limitations

Our ablations are in-corpus: Task 2 gates use the same 253 public examples for development and evaluation, and Task 3 variants use the same 76 labeled training examples. Only the final workflows were submitted to the leaderboards, limiting stage-level conclusions out of sample. The end-to-end workflows also use more model calls than direct prompting, so these comparisons cannot disentangle the contribution of validation gating from that of additional inference compute. Controlled updates may still admit harmful proposals or reject useful ones. For Task 3, ROUGE-1 does not fully capture attribution or numerical faithfulness, and the overall F1 gain is small. API-hosted inference is also stochastic and provider-dependent, and we do not estimate run-to-run variance.

References

- Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2024. [AgentVerse: Facilitating Multi-Agent collaboration and exploring emergent behaviors](#). In *International Conference on Learning Representations*.
- Zhiyu Chen, Wenhui Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. [FinQA: A dataset of numerical reasoning over financial data](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711. Association for Computational Linguistics.
- Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. [Chain-of-Verification reduces hallucination in large language models](#). *arXiv preprint arXiv:2309.11495*.
- Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. [Enabling large language models to generate text with citations](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488. Association for Computational Linguistics.
- Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujia Yang, Minlie Huang, Nan Duan, and Weizhu Chen. 2024. [CRITIC: Large language models can Self-Correct with Tool-Interactive critiquing](#). In *International Conference on Learning Representations*.
- Aaron Grattafiori et al. 2024. [The Llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. [Self-Refine: Iterative refinement with Self-Feedback](#). In *Advances in Neural Information Processing Systems*, volume 36.
- Xueqing Peng, Lingfei Qian, Yan Wang, Ruoyu Xiang, Yueru He, Yang Ren, Mingyang Jiang, Vincent Jim Zhang, Yuqing Guo, Jeff Zhao, Huan He, Yi Han, Yun Feng, Yuechen Jiang, Yupeng Cao, Haohang Li, Yangyang Yu, Xiaoyu Wang, Penglei Gao, and 28 others. 2026a. [MultiFinBen: Benchmarking large language models for multilingual and multimodal financial application](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16934–16963, San Diego, California, United States. Association for Computational Linguistics.
- Xueqing Peng, Ruoyu Xiang, Fan Zhang, Mingzi Song, Mingyang Jiang, Yan Wang, Lingfei Qian, Taiki Hara, Yuqing Guo, Jimin Huang, Junichi Tsujii, and Sophia Ananiadou. 2026b. [Ebisu: Benchmarking large language models in Japanese finance](#). *arXiv preprint arXiv:2602.01479*.
- Qwen Team. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. [ReAct: Synergizing reasoning and acting in language models](#). In *International Conference on Learning Representations*.
- Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. [TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287. Association for Computational Linguistics.

A Workflow and Reproducibility Details

This appendix summarizes the transition scopes used by the Task 2 resolver, the evidence-state controls used for Task 3, and additional ablation results. The model assignments and evaluation settings are reported in Section 4.

A.1 Task 2 Transition Control

The resolver maintains a single incumbent label. Gated composition first combines the rubric and decomposition candidates using direction and total-withholding checks. Subsequent diagnostics operate in sequence on commitment strength, polarity

and relation, and the final adjacent-label boundary. Each diagnostic may propose only its predefined transition. If a proposal is invalid, unsupported, or outside that scope, the incumbent is unchanged. For an incumbent $y^{(k)}$, proposal $\hat{y}^{(k)}$, and admission decision $a_k \in \{0, 1\}$, the update is

$$y^{(k+1)} = \begin{cases} \hat{y}^{(k)}, & a_k = 1, \\ y^{(k)}, & a_k = 0. \end{cases} \quad (1)$$

Table 2 summarizes the transition scope and admission logic exposed to each diagnostic. The resolver never asks a later diagnostic to produce a new unrestricted five-way prediction.

Check	Transition scope and admission rule
Direction	Triggered by broad-direction disagreement between the initial candidates. A transition is admitted only when two valid direction records agree, match the direction of an existing candidate, and do not identify the response as multi-part.
Total with-holding	Triggered when refusal language may apply to all or only part of the request. A refusal-direction transition is admitted only when the evidence indicates that the full requested proposition is withheld rather than only a secondary detail.
Commitment strength	Applied after a positive direction has been retained but its strength remains ambiguous. Only an adjacent $+1 \leftrightarrow +2$ transition is allowed, and the supporting record must justify the boundary change without reversing direction.
Polarity and relation	Triggered when a parsed polarity or question-response relation conflicts with the incumbent. Only the predefined repair for that diagnostic is allowed; the record must be valid and support the declared relation.
Adjacent boundary	Applied when the final incumbent lies near a label boundary. The audit may propose only one label-local transition, which must be valid, adjacent to the incumbent, and consistent with the audited boundary.

Table 2: Task 2 transition checks. A failed admission check leaves the incumbent label unchanged.

A.2 Task 2 Boundary-Audit Outcomes

The final boundary audit proposed 75 transitions relative to the incumbent produced after the preceding repairs. Table 3 separates the realized effect of admitted proposals from the counterfactual effect of rejected proposals. Applying every proposal would have produced a net loss of two correct predictions; selective admission instead produced a net gain of 11.

Decision	Total	Corr.	Harm.	Neutral	Net
Admitted	19	13	2	4	+11
Rejected	56	17	30	9	-13
All	75	30	32	13	-2

Table 3: Task 2 boundary-transition outcomes. Net is corrective minus harmful; the rejected row is counterfactual.

A.3 Task 3 Evidence State and Validation

Task 3 represents answer requirements and candidate findings separately so that retrieval alone does not authorize synthesis. The logical evidence state contains the following record types; these describe information passed between modules rather than requiring a particular serialization format.

Answer requirement. The requested answer object and relation, assigned source responsibility, and source dependencies define what must be supported.

News finding. The record carries company attribution, the supported action or relation, the relevant passage, and source provenance.

Financial-statement evidence. The record identifies the metric, statement type, period, unit, value, required calculation, and provenance.

Contextual or unresolved material. Topically relevant context is not treated as answer support, and a requirement without an admitted record remains explicitly unresolved.

News findings are validated for company attribution, answer-object relevance, and support for the relation requested by the question. Financial-statement records are checked for compatible metric, statement type, period, unit, and calculation. Deterministic matching and arithmetic are used where possible; semantic interpretation is reserved for unresolved field meaning. Multilingual evidence may be normalized into English only after validation, while preserving the entity, action, number, period, and supported relation. The synthesizer receives admitted records and applies coverage, attribution, numerical-grounding, and unsupported-claim checks. If restricted revision cannot safely repair a defect, the affected requirement remains unresolved or retains evidence-derived wording.

A.4 Task 3 Section-Level Results

Table 4 separates the Answer and Evidence sections used by the Task 3 output format. The structured variants improve Evidence F1 substantially, while their Answer F1 remains below direct prompting,

showing that improved evidence state does not automatically yield better open-ended synthesis.

System	Section	P	R	F1
Direct prompting	Overall	0.3886	0.3437	0.3399
	Answer	0.4938	0.2548	0.2996
	Evidence	0.1703	0.1647	0.1570
Filtered Financial Evidence	Overall	0.5202	0.3070	0.3415
	Answer	0.4090	0.2638	0.2894
	Evidence	0.5139	0.5134	0.5135
Final workflow	Overall	0.4600	0.3269	0.3425
	Answer	0.3366	0.2917	0.2893
	Evidence	0.5132	0.5132	0.5132

Table 4: Task 3 ROUGE-1 ablations on 76 training examples. Scores are macro-averaged with stemming disabled; bold marks the best value in each column within each section.

A.5 Reproducibility Note

The workflow stages and deterministic gate logic are described above, and gold labels or reference answers are used only for evaluation. Because the systems rely on hosted API models and complete decoding metadata were not archived for every run, exact regeneration of intermediate model outputs is not guaranteed.