

Daikoku at FinEval 2026: Multi-Stage Labelling with Forced-Choice Adjudication

Kyosuke Morita

Lloyds Banking Group

Kyosuke.Morita@lloydsbanking.com

Abstract

This paper describes our submission to Sub-Task 2 of the FinEval 2026 Multilingual Financial Language Understanding Challenge, Japanese Financial Implicit Communication Recognition (JF-ICR), which places a company's question-answer on a five-point commitment scale. Our submitted system, a three-stage pipeline, scores 15/15 public and 33/35 private, 4th of 10. A single open-weight model labels each of the 50 test items independently by verbaliser scoring; that labelling is audited against the class distribution the organisers publish in advance, which yields a label-free lower bound F on its error count and names the classes those errors sit in; each over-full class is then handed back to the model as a forced choice over its own members, with the exact number of mislabelled items asserted. The audit opens at $F=3$ and closes at $F=0$. Enforcing the same counts directly, by min-cost assignment, also reaches $F=0$ but recovers only the errors already nearest the decision boundary. The official score bounds what a floor of zero is worth: two of the 50 items are wrong beneath it, and, both the submitted labelling and the gold being balanced, those two errors can only form a compensating transposition.

1 Introduction

Japanese financial communication states commitment indirectly. Modality and negation occur at the absolute end of the sentence, refusals are hedged rather than asserted, and disclosure convention discourages a plain "no" (Peng et al., 2026b). The Japanese Financial Implicit Communication Recognition (JF-ICR) task makes that judgement explicit: given a Japanese financial question–response, place the response on a five-point scale running from strong commitment at +2 to outright refusal at −2.

EBISU (Peng et al., 2026b), from which the benchmark for this task is drawn, evaluates twenty-two systems on JF-ICR spanning general-purpose,

Japanese-adapted, English-financial and Japanese-financial families. The best reaches 0.606 accuracy and the strongest proprietary system 0.553, and EBISU reports that scale buys only limited gains while language- and domain-specific adaptation does not reliably help: the gap is not one that capacity or in-language pretraining is closing on its own.

Our submission has three stages:¹ (i) an LLM labels each item independently by verbaliser scoring; (ii) that labelling is audited against the class distribution quota; (iii) each over-full bucket is handed back to the model as a forced choice.

Any known marginal splits the errors of a labelling into two disjoint parts and detects exactly one of them — it is an *exhaustive* detector for errors that change the class counts, naming the buckets they sit in, and *blind* to permutations of labels among items that leave every count intact. Stages 2 and 3 exploit the exhaustiveness, and exploiting it needs a second opinion rather than a better decision rule: enforcing the counts directly by min-cost assignment recovers only the errors that already sit nearest the decision boundary. The official score of the submitted file demonstrates the blindness.

2 Related Work

Sentiment and stance classification over financial text has moved from supervised models trained on annotated news and analyst reports (Malo et al., 2014; Araci, 2019) to domain-adapted and instruction-following LLMs evaluated on broad benchmark suites (Wu et al., 2023; Xie et al., 2024). Coverage outside English is both thinner and measurably harder: frontier models lose a large part of their headline accuracy once the language is not English (Peng et al., 2026a), and Japanese-specific benchmarks report the same gap (Hirano, 2024;

¹Codebase is available at <https://github.com/kyosek/JF-ICR-Daikoku-Solution>.

Peng et al., 2026b). JF-ICR sits in exactly that setting, with 253 labelled items in the open split, which we therefore approach as a prompting problem rather than a training one. Constraining structured predictions by global count statistics is in turn an old device (Roth and Yih, 2004; Chang et al., 2007; Ganchev et al., 2010); what a shared task adds is that the constraint is published, exact, and free.

3 Challenge Dataset and Setup

The organiser prepares two data splits (Table 1): an open split of 253 instances with released gold labels, and a test split of 50 instances whose labels are withheld and which is the submission target. The test split is further divided into 15 public and 35 private items (3 and 7 per label). During the evaluation phase an uploaded file is scored on the public items only, so a configuration can be given an aggregate public score before the deadline; but the organisers release neither the split column nor the test gold, so that score cannot be attributed to individual items, and a private score exists only for the final submission. The two splits differ in three aspects:

(i) **Label distribution.** The open split is heavily skewed: +1 alone accounts for 54.9% of instances and −2 occurs twice. The test split is published as exactly balanced, 10 instances per label — the constraint that stages 2 and 3 of §4 utilise.

(ii) **Length.** Open instances are roughly two and a half times longer, 502 characters of question–response block against 200.

(iii) **Multi-part questions.** The detector of Appendix E flags 76 of the 253 open-split questions as asking more than one thing — a manual audit there confirms 52 — and none at all on the test split, every one of whose questions is a single sentence.

As the gold labels of the test split are withheld, the author annotated all 50 test items after the deadline from the EBISU guidelines (Peng et al., 2026b) and the question–response alone (Appendix B). We report agreement with that pass as a sanity check and treat it throughout as a weak non-independent reference.

The official metric is accuracy, which we report throughout, together with macro- F_1 (Sokolova and Lapalme, 2009) on the open split, whose skewed label distribution accuracy alone does not discriminate.

	Open	Test
n	253	50
+2	29 (11.5%)	10 (20%)
+1	139 (54.9%)	10 (20%)
0	72 (28.5%)	10 (20%)
−1	11 (4.3%)	10 (20%)
−2	2 (0.8%)	10 (20%)
Mean Q–R length (chars)	502	200
Single-sentence questions	121 (47.8%)	50 (100%)
Multi-part (detector)	76 (30.0%)	0 (0%)
audited	52 (20.6%)	0 (0%)

Table 1: Split statistics. Open-split counts are the released gold labels; the test-split distribution is the balanced design published by the organisers. “Multi-part” is the detector of Appendix E; “audited” is the subset of its flags confirmed by hand there.

Algorithm 1: The submitted pipeline

```

input : items  $x_{1:n}$ ; published class quota  $q$ 
output : labelling  $\hat{y}_{1:n}$  meeting  $q$ 

// Stage 1: independent verbaliser scoring
1 for  $i \leftarrow 1$  to  $n$  do
2    $P(y | x_i) \leftarrow \text{softmax}_{y \in \mathcal{Y}} \log p_\theta(v_y | x_i)$ 
3    $\hat{y}_i \leftarrow \arg \max_y P(y | x_i)$ 

// Stage 2: audit against the quota
4  $n_y \leftarrow |\{i : \hat{y}_i = y\}|$    for all  $y \in \mathcal{Y}$ 
5  $F \leftarrow \sum_y \max(0, n_y - q_y)$    // error floor
6  $O \leftarrow \{y : n_y > q_y\}; \quad U \leftarrow \{y : n_y < q_y\}$ 

// Stage 3: adjudicate each over-full
//          bucket, then re-audit
7 foreach  $y \in O$  do
8    $B \leftarrow \{i : \hat{y}_i = y\}; \quad s \leftarrow n_y - q_y$ 
9    $(\pi, d) \leftarrow \text{FORCEDCHOICE}(B, y, U, s)$ 
10  // §4.3
11  foreach  $i \in \pi_{1:s}$  do
12     $\hat{y}_i \leftarrow d_i \in U$    // voted destination

12  $n_y \leftarrow |\{i : \hat{y}_i = y\}|$    for all  $y \in \mathcal{Y}$    // re-audit
13 if  $\sum_y \max(0, n_y - q_y) \neq 0$  then abort
14 else write a submission file

```

4 Submitted System

Our submission is the three-stage pipeline of Algorithm 1. Stage 1 labels each item independently; stages 2 and 3 then exploit a constraint that holds over the labelling *jointly* and that no per-item method can see — that the test split’s class distribution is published in advance.

4.1 Stage 1: Verbaliser Scoring

We splice the EBISU annotation guidelines (Peng et al., 2026b) — the per-label linguistic signals and representative examples the human annotators worked from — immediately before the question–response block. We also supply a scope condition the guidelines leave open: they state both that the

label reflects the *strongest applicable* commitment and that annotation follows the *overall stance of the full response*, and the two diverge when one response carries spans at different commitment levels. We resolve this by scope rather than by preference, stating the most-committal directive only for *multi-part* questions, where the response is a set of sub-answers and the item has no determinate label until the prompt says which part governs. The gate is the detector of Appendix E, which also shows what stating the rule unconditionally costs.

For each item we render the chat-templated prompt, tokenise each of the five label strings $v_y \in \{+2, +1, 0, -1, -2\}$, and teacher-force through the label’s tokens, summing token log-probabilities to obtain $\log p_\theta(v_y | x)$. A softmax over the five gives $P(y | x)$, and we take the argmax.

4.2 Stage 2: The Quota as a Lower Bound

The organisers publish the test split as balanced: $q_y = 10$ for each of the five labels across the 50 items. If a labelling predicts class y exactly n_y times, it is wrong on at least

$$F = \sum_{y \in \mathcal{Y}} \max(0, n_y - q_y) \quad (1)$$

items. This is a certificate rather than an estimate: no labelling with those counts can do better, and F is computable before any label is seen. The labelling we evaluate it on is stage 1’s: $n_y = |\{i : \hat{y}_i = y\}|$ counts the argmax predictions of §4.1, before any reassignment. The over-full classes O name the buckets those surplus errors live in, shrinking the audit surface to their union, and the under-full set U enumerates the labels the displaced items can move to. The quota does not say *which* item in a bucket is wrong, so it is a *detector* rather than a *corrector* and finding the items still needs a second opinion.

4.3 Stage 3: Forced Choice

An over-full bucket holds the n_y items stage 1 assigned to y , of which exactly $s = n_y - q_y$ are mislabelled, so we instruct the model to compare them: *These N items were all labelled +2. Exactly s of them are mislabelled; the true label of those items is 0 or -1. Which?*

FORCEDCHOICE takes the bucket B , its class y , the under-full set U and the surplus s , and returns an ordering π of B from most to least likely mislabelled together with a destination $d_i \in U$ for each of its first s entries, which are the entries reassigned.

Because it is a forced choice over a shortlist rather than open five-way classification, the candidates calibrate each other instead of each being judged against an internalised threshold: a model that is confidently wrong item-by-item can still rank the odd one out, since ranking does not require the absolute threshold that classification gets wrong. A single call cannot rule out position bias (Wang et al., 2024; Tang et al., 2024), so we ask $r=5$ times under rotated candidate orderings — rotation rather than shuffling, to stay reproducible without a seed — and vote over each round’s top s picks. The stage closes by recomputing Eq. 1 on the re-assigned labelling and writing the submission file only if $F = 0$.

When $|U| = 1$ the destination is forced by arithmetic, and the model is asked only which items to move. When $|U| > 1$ the model’s own $P(y | x)$ cannot settle it. On precisely the items stage 1 has got wrong, those posteriors are the miscalibrated ones, so reading a destination off them would reintroduce the absolute threshold the forced choice exists to avoid. The prompt names the under-full labels and requires each pick in item:label form, and the destination is settled by the same $r=5$ vote that settles the items, counted over those rounds in which the item was among the round’s top s . Ties in either vote go to the candidate named earliest in the rotation sequence, which is deterministic given a stored stage 1 output; a round naming fewer than s items contributes the picks it does make, and a label outside U is discarded. Buckets are adjudicated independently against the stage 1 labelling and their reassignments applied together, so the result does not depend on the order in which O is traversed. Nothing in the mechanism stops the model sending more items to an under-full class than that class is short of, and nothing repairs it if it does. The re-audit catches an infeasible assignment, and the pipeline aborts rather than write a file.

4.4 Robustness to an Approximate Quota

For this challenge, the organiser released the class distribution and stages 2 and 3 exploit it. In the real world, often the class distribution is not known exactly, and the published marginal may be an estimate or a historical carry-over. In such cases, the mechanism degrades linearly. Write q for the true class counts and $\tilde{q}$ for the published ones, both

totalling n , and let

$$\Delta = \frac{1}{2} \|\tilde{q} - q\|_1 = n \cdot d_{\text{TV}}\left(\frac{\tilde{q}}{n}, \frac{q}{n}\right), \quad (2)$$

where d_{TV} is total variation distance — half the L_1 distance between two distributions on the same support — so Δ is that distance stated in items: the number of labels separating the published marginal from the true one. Because each summand $q_y \mapsto \max(0, n_y - q_y)$ is non-increasing and 1-Lipschitz in q_y , and because q and $\tilde{q}$ share a total,

$$|F(\tilde{q}) - F(q)| \leq \Delta. \quad (3)$$

So $F(\tilde{q}) - \Delta$ remains a valid certificate: a marginal known to be accurate to within Δ items buys a floor Δ items weaker, and nothing further about q need be known. Stage 3 moves exactly $F(\tilde{q})$ items, so at most Δ of those moves are in excess of what the true marginal licenses; the exposure scales with the error in the published marginal. Neither statement needs the split to be balanced, and both hold for any published marginal.

Enumerating every non-negative integer marginal at each distance $\Delta \leq 6$ from the released open-split gold counts, Eq. 3 holds in all 4,395 cases and is met with equality at every Δ , so it cannot be tightened. The median deviation is exactly $\lceil \Delta/2 \rceil$, so it is also pessimistic for a marginal that is noisy rather than adversarial. The practical margin is nevertheless narrow: an audit that opens at F tolerates a published marginal in error by at most $F - 1$ items before the floor it certifies falls to zero, and the stage 1 audits reported in §5.1 open at values of F in the low single digits.

4.5 Reproducibility

The submitted system uses one model, Qwen3.6-35B-A3B (Qwen Team, 2026a), served locally from a 22 GB GGUF quantisation (UD-Q4_K_M) with reasoning mode off, on one Apple M3 Max with 48 GB of unified memory and greedy decoding (temperature = 0). An end-to-end run over the test split takes just over four minutes. Greedy decoding is nevertheless *not* bit-exact under llama.cpp: repeated runs of an identical configuration differ on one or two items, so stages 2 and 3 are reproducible only over a stored stage 1 output, which we release with the code. Alternative models are compared in §5.3, and the full setup is in Appendix A.

Pipeline prefix	F	Agr.	Official Pub.	Pri.
Stage 1, submitted run	3	47	14/15	—
+ 2&3, submitted run	0	50	15/15	33/35
<i>Alternative uses of the same quota</i>				
Min-cost assignment	0	48	—	—
Ensemble majority vote	1	49	—	—
Consensus-spread ranking	0	50	—	—

Table 2: Cumulative stage ablation on the test split. F is the provable error floor of Eq. 1; “Agr.” is agreement out of 50 with our post-hoc annotation; “Official” is the organisers’ score. Only two files were uploaded — the raw stage 1 argmax during the evaluation phase and the submitted file.

5 Results

5.1 Results Overview

The submitted file scores 15/15 public and 33/35 private, fourth of ten and tied with two other systems. The leaderboard is tight: the nine non-outlier systems span 34/35 to 30/35, none reached 35/35, and the first-placed systems are one item ahead of ours.

Table 2 reports the performance of the pipeline as we add one pipeline stage at a time on the test split. The stage 1 run behind the submitted file over-predicts +2 once and labels items 259 and 296 as +1 rather than -1 ($F=3$, audit surface 23 items). Stages 2 and 3 move all three: $F=0$, 50/50 agreement, 15/15 public and 33/35 private, and that labelling is the submitted file. The pipeline is therefore more stable than the stage it is built on, which is the practical case for it.

Only stage 3 moves a label: stage 2 changes nothing and emits F together with the over-full buckets. With $r=5$ rotations its votes are decisive (id=300 named first in all five orderings of the +2 bucket, id=296 in all five of the +1 bucket, id=259 in three of five). Both buckets leave $U = \{0, -1\}$, so this run exercises the $|U| > 1$ path and the destination was the model’s to choose in each case and it was voted unanimously. $300 \rightarrow 0$ and $296 \rightarrow -1$ in every rotation, $259 \rightarrow -1$ in all three rotations that picked it, so the tie-break never fired. The three destinations happen to meet the one-item deficit at 0 and the two-item deficit at -1 exactly, and the re-audit passed without repair — but that is an observation about this run, not a property the mechanism enforces.

The lower block of Table 2 is the evidence for treating the quota as a detector rather than a cor-

rector. Enforcing it directly — a min-cost assignment over $\log P(y \mid x)$ for the highest-probability labelling that meets the counts — reaches $F=0$ but stops at 48/50 from either stage 1 run, two items short of the forced choice. Min-cost buys its counts as cheaply as possible, so it flips the smallest-margin items in each bucket. That is the right item only where the rest of the bucket is saturated, as in the +2 bucket; in the +1 bucket the second real error is merely the third cheapest, so min-cost breaks a correct item and never reaches it, while forced choice names it in all five rotations (Appendix C). Cheapest-to-flip and most-likely-wrong coincide when the bucket is saturated and come apart when it is not; a comparative re-ask reaches the second regime, and no re-weighting of the same posterior can.

A majority vote over 16 test-split runs reproduces the submitted stage 1 argmax on all 50 items, so ensembling by vote does not improve upon it. Ranking each flagged bucket by how far those runs *disagree* instead recovers the same three items as stage 3, at no extra inference cost wherever an ensemble already exists (Appendix C).

5.2 Implications for the Forced-choice Stage

The submitted pipeline carries $F=0$ and is wrong on two items. It predicts each class exactly 10 times and the gold is balanced too, so two labellings that differ on exactly two items must differ by a transposition: a pair (i, j) with $G(i) = S(j)$ and $G(j) = S(i)$. That is precisely the compensating pair Eq. 1 cannot see, so the floor was correct and uninformative at once — the blind half of §1, measured on a submitted file rather than asserted.

Our post-hoc annotation agrees with the submitted file on all 50 items (Appendix B), so it carries the same class counts and Eq. 1 certifies it at $F=0$ as well: the floor is a property of the counts, not of the labelling that realises them. This also fixes the reading of the “Agr.” column, which orders configurations but saturates at 50/50 with two errors still in the file.

5.3 Open Split

We used the open split, where gold is released, to fix the prompt and choose the model; Appendix D reports it in full. Qwen3.6-35B-A3B is the strongest of eight models and was submitted at 0.676 accuracy against 0.557 for the best of four non-LLM baselines — a 13-point gap that neither cue-phrases counts nor a 110M encoder fine-tune

closes. The two splits are not commensurable: a constant predictor scores 0.549 on the open split against 0.200 on the test split, so the three data differences of §3 account for most of the gap before any modelling question arises.

6 Conclusion

We submitted a three-stage pipeline that labels each JF-ICR item independently and audits the labelling against the known class distribution. The forced-choice mechanism demonstrates that a model that is confidently wrong item-by-item can still rank the odd one out, because ranking does not require the absolute threshold that item-by-item classification gets wrong.

Limitations

Test gold was withheld, so the only externally scored figures in this paper are the leaderboard’s: 15/15 and 33/35 for the submitted file, and the 14/15 public-only score returned for the raw stage 1 argmax during the evaluation phase. Every other test-split figure is agreement with a single-annotator pass written by the author. It is not independent, and it coincides with the submitted file on all 50 items, which the official score shows is wrong on two: agreement with it is therefore consistent with two errors and saturates before it can bound them. We use it only to order configurations — no claim rests on a one-item difference in it, and the ablation is led by the label-free floor F for this reason — and neither it nor the floor locates the two errors.

A compensating pair of errors leaves every class count untouched, so $F=0$ means consistent-with-correct class distribution rather than correct: the submitted file has $F=0$ and is wrong on two items. The mechanism also inherits the known class distribution: by Eq. 3 a marginal in error by Δ items costs Δ items of certificate and licenses at most Δ moves stage 3 should not have made. Since the audits open in the low single digits, that tolerance is a few items wide and no more, and we have not measured what a wrong surplus does to the forced choice itself.

Acknowledgements

We thank the organisers for running the shared task and we thank the reviewers for their careful reading and constructive feedback.

References

- Anthropic. 2026. [Claude Opus 5 system card](#).
- Dogu Araci. 2019. [FinBERT: Financial sentiment analysis with pre-trained language models](#). *ArXiv*, abs/1908.10063.
- Ming-Wei Chang, Lev Ratinov, and Dan Roth. 2007. [Guiding semi-supervision with constraint-driven learning](#). In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 280–287, Prague, Czech Republic. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Kuzman Ganchev, João Graça, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11:2001–2049.
- Gemma Team. 2026. [Gemma 4 technical report](#). *arXiv preprint arXiv:2607.02770*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. [DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning](#). *Nature*, 645(8081):633–638.
- Masanori Hirano. 2024. [Construction of a Japanese financial benchmark for large language models](#). In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing*, pages 1–9, Torino, Italia. Association for Computational Linguistics.
- Leonard Lin and Adam Lensenmayer. 2025. [Shisa V2: A family of bilingual Japanese–English models](#). Shisa.AI, Hugging Face model card.
- Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Walenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.
- OpenAI. 2025. [gpt-oss-120b & gpt-oss-20b model card](#). *arXiv preprint arXiv:2508.10925*.
- 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, et al. 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, et al. 2026b. [Ebisu: Benchmarking large language models in japanese finance](#). *arXiv preprint arXiv:2602.01479*.
- Qwen Team. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- Qwen Team. 2026a. [Qwen3.6-35B-A3B: Agentic coding power, now open to all](#).
- Qwen Team. 2026b. [Qwen3.8: A new bar for coding and cowork](#).
- Dan Roth and Wen-tau Yih. 2004. [A linear programming formulation for global inference in natural language tasks](#). In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, pages 1–8, Boston, Massachusetts, USA. Association for Computational Linguistics.
- Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437.
- Raphael Tang, Crystina Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Ture. 2024. [Found in the middle: Permutation self-consistency improves listwise ranking in large language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2327–2340, Mexico City, Mexico. Association for Computational Linguistics.
- Tohoku NLP Group. 2023. [BERT base Japanese \(unidic-lite with whole word masking, CC-100 and jawiki-20230102\)](#). Hugging Face model card.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. [Large language models are not fair evaluators](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.
- Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhjanj Kam-badur, David Stuart Rosenberg, and Gideon Mann. 2023. [BloombergGPT: A large language model for finance](#). *ArXiv*, abs/2303.17564.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyang Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, et al. 2024. [Finben: A holistic financial benchmark for large language models](#). In *Advances in neural information processing systems*, volume 37, pages 95716–95743.

A Pipeline and Inference Setup

Inference setup and cost. The model is served locally from a 22 GB GGUF quantisation (UD-Q4_K_M) with reasoning mode off, on one Apple M3 Max with 48 GB of unified memory, under a native arm64 build with a context window of 8192 and all layers offloaded to the GPU. Decoding is greedy (temperature 0) with a 32-token generation budget. A full stage 1 pass over the open split (253 items) costs 12 min 48 s of server time (3.0 s per instance); over the 50-item test split it costs 3 min 12 s, and an end-to-end run of the pipeline over that split — server start-up, stage 1, both audits and the ten adjudication calls — takes just over four minutes.

B Post-hoc Test-Split Annotation

Protocol. One author labelled all 50 test items after the submission deadline, in dataset order, from the official annotation guidelines and the Japanese question–response alone. No system prediction, posterior or class count was visible during the pass, and the published balance constraint was deliberately *not* enforced. That the resulting counts satisfy it anyway is an observation about the pass, not a constraint imposed on it.

Counts. +2: 10, +1: 10, 0: 10, −1: 10, −2: 10. Applying Eq. 1 to this labelling gives $F = 0$: no class is over quota, so the annotation is certified by exactly the same test that certifies the submitted file — and, as §5.2 records, on this split that certificate is compatible with two errors.

Relation to the submitted file. The annotation and the submitted file coincide on all 50 items, so the only disagreements to report are with stage 1. They are the three items below, which are precisely the ones stage 3 relabels (§4.3); each moves onto the label the annotation gives it.

Only 259 is near the stage 1 decision boundary; on the other two the per-item posterior is confidently on the label the annotation rejects. On 300 that still costs nothing, because the rest of its bucket is

id	stage 1	annotation	$P(\text{stage 1})$	$P(\text{annot.})$
259	+1	−1	0.506	0.452
296	+1	−1	0.829	0.152
300	+2	0	0.888	0.007

saturated at 0.998 or above and confidence alone isolates it; on 296 it does, and 296 is precisely the item min-cost assignment misses and forced choice recovers (§5.1). The coincidence on all 50 items is not evidence that either labelling is correct: one annotator reading the same guidelines that the prompt was written from can converge on the same labelling including its errors, and the official score (§5.2) shows that two of the 50 are wrong in both.

C Alternative Uses of the Quota

This appendix expands the lower block of Table 2. All three alternatives consume the same published quota as stages 2 and 3; none of them beats the forced choice.

Min-cost assignment. Within each over-full bucket the assignment flips the s items of smallest margin, which is the right rule only when the error is also the least confident candidate. In the rerun’s +2 bucket it is: the surplus is 1 and the real error, id 300, is the least confident of the eleven at $P=0.888$ against 0.998 or above for the other ten. In the +1 bucket it is not: the surplus is 2 and the two real errors rank first (id 259, 0.506) and third (id 296, 0.829), so the assignment takes the two cheapest — correcting 259, breaking 288 (0.584) and never reaching 296. Stage 3 inverts exactly that pair, naming 296 in all five rotations and 288 in one.

Ensemble majority vote. The per-item majority over 16 full test-split runs, spanning four models under varying prompt and decoding settings. It reproduces the submitted stage 1 argmax on all 50 items, the wrong one included: on id 300 the runs split 12 +2 to 3 0 and 1 −1, so the ensemble does not improve upon the argmax and the vote carries no information the argmax did not already have.

Consensus-spread ranking. The same 16 runs read for their disagreement rather than their majority: inside each quota-flagged bucket, candidates are ranked by how many runs agree with the flagged label and the least-agreed s are reassigned. In the +2 bucket ten of the eleven candidates are unanimous at 16/16 and id 300 is the sole exception

at 12/16. It selects the same three items with the same destinations as stage 3. A wrong prediction every run agrees on stays invisible to it, so it is a cheaper substitute for stage 3 only where an ensemble already exists.

D Open Split Results

Systems. The learned non-LLM baselines are a logistic regression over counts of hand-curated Japanese hedging and commitment cue phrases, and a fine-tune of bert-base-japanese-v3 (Tohoku NLP Group, 2023) — a 110M-parameter BERT (Devlin et al., 2019) pre-trained on Japanese Wikipedia and the Japanese portion of CC-100 — run once with the default cross-entropy objective and once with class-balanced loss. The models compared are Qwen3.6-35B-A3B (Qwen Team, 2026a), Qwen3.8-27B (Qwen Team, 2026b), Qwen3-30B-A3B-Instruct-2507 (Qwen Team, 2025), Gemma-4-12B-it (Gemma Team, 2026), GPT-OSS-20B (OpenAI, 2025), DeepSeek-R1-Distill-Qwen-32B (Guo et al., 2025) and Shisa-v2-Qwen2.5-32B (Lin and Lensenmayer, 2025), all served locally as GGUF quantisations under the setup of §4.5, together with Claude Opus 5 (Anthropic, 2026), which is called through the vendor API. Table 3 abbreviates these names.

Table 3 reports the open split, where gold labels are released. The prompt scores 0.684 accuracy under free generation and 0.676 under the verbaliser scoring the submission uses, so the decoding interface is close to inconsequential for accuracy and the residual error does not live in label extraction. All three out of four non-LLM baselines; the cuelexicon regression and the unweighted encoder sit within one point of a constant +1 predictor on accuracy while differing by more than ten points in macro-F₁. The unweighted encoder collapses onto the majority class (macro-F₁ 0.153) and accuracy alone does not show it. The gap from the best baseline to the LLM is 13 points of accuracy and 16 of macro-F₁, so the label is not recoverable from cuephrase counts or from fine-tuning a 110M encoder on ≈200 examples per fold, and the LLM comparison is measuring something other than surface form.

Among models, Qwen3.6-35B-A3B is the strongest on accuracy and was submitted on that basis; Claude Opus 5 is 2.0 points behind on accuracy but 6.7 ahead on macro-F₁, the only model appreciably better than chance on the two rare classes.

System	Acc.	M-F ₁
<i>Non-LLM baselines (5-fold CV, out-of-fold)</i>		
Majority class (+1)	0.549	0.142
Cue lexicon + LR	0.557	0.251
Encoder fine-tune	0.553	0.153
class-balanced	0.522	0.249
<i>LLMs, zero-shot, identical prompt, greedy decoding</i>		
Qwen3.6-35b-a3b (ours)	0.684	0.407
verbaliser (stage 1)	0.676	0.404
Claude Opus 5 (API)	0.664	0.474
Qwen3.8-27b	0.652	0.382
Gemma-4-12b-it	0.644	0.363
Qwen3-30b-a3b	0.605	0.335
GPT-OSS-20b	0.601	0.352
Shisa-v2-Qwen2.5-32b	0.534	0.292
DeepSeek-R1-Distill-32b	0.486	0.303

Table 3: Open split ($n=253$, released gold), accuracy and macro-F₁. Learned baselines are out-of-fold predictions from 5-fold CV; stratified CV is infeasible because the −2 class has two members.

The Japanese-specialised fine-tune is second to last, so in-language specialisation at 32B does not substitute for general capability.

E Multi-Part Aggregation

Two guideline clauses in tension. The aggregation rule is not an addition to the official guidelines; it resolves an ambiguity already in them. EBISU’s general annotation rules (Peng et al., 2026b, Appendix A.2) contain both of the following:

When multiple signals appear, the label reflects the strongest applicable commitment or refusal expressed in the answer.

Annotation is based on the overall stance of the full response, rather than isolated sentences or individual expressions; labels reflect the speaker’s dominant commitment or refusal as conveyed by the answer as a whole.

Read as unrestricted rules over a response, the two disagree on exactly the responses that matter. Where a response carries spans at different commitment levels, the first clause takes the maximum over those spans and the second takes the dominant stance across them: a hedged answer containing one committal sentence is +1 under the first rule and 0 under the second. Table 4 is that disagreement in the data. Item (a) is labelled correctly only under the strongest-signal clause — its two asks are answered at +1 and 0, and reading the response as a whole yields the 0 that the gold label

rejects. Item (b) is labelled correctly only under the overall-stance clause — one ask, one long hedged response, and the strongest-signal reading latches onto the committal closing span and returns the +1 that the gold label rejects.

Resolution by scope. The clauses are not contradictory once the quantifier they are silent about is supplied. A multi-part question makes the response a set of sub-answers, one per ask; a single-part question makes it one answer. Strongest-signal is then the rule *across* asks and overall-stance the rule *within* one, and both items of Table 4 are recovered. This is why the directive in the stage 1 prompt (§4.1) is gated rather than stated unconditionally. EBISU never has to state the scope condition, because its curation removes multi-part items before annotation begins; the released open split retains them (§3), so a participant prompting from those guidelines has to decide the case that the guidelines do not.

Detector. The gate is a lexical heuristic over the question text alone. It splits the question on sentence-final punctuation and line breaks, counts the sentences carrying a request or interrogative marker (〜ください, 〜てほしい, 〜伺いたい, a trailing か/の before the full stop, or an interrogative word), and flags the question when two or more sentences ask something, or when an additive connective (また、, 加えて, さらに, もう一点, 2点目) occurs in a question of more than one sentence. Keying on the question particle か alone would miss most of them: Japanese investor-relations questions are usually phrased as requests rather than as syntactic interrogatives.

Validation. We audited all 76 flagged open-split questions by hand — a single-annotator pass by the author, like the annotation of Appendix B, whose item-level judgements accompany our released code — counting an item as multi-part when it poses two or more asks that could be answered at different commitment levels, and not when the second ask restates the first, offers an A-or-B alternative within one ask, or is preamble. 52 of 76 are confirmed, a precision of 0.68. The 24 rejections are one ask restated or elaborated (14), a single ask split by a line break inherited from the source disclosure (5), an A-or-B alternative within one ask (3), 合わせて matched inside the non-additive 〜に合わせて (1), and one item whose question and response fields are transposed in the release

(1). A sample of 30 of the 177 unflagged questions contains two genuine multi-part items, both posing their asks in nominal form (〜は。), which carries no marker the detector keys on; extrapolated, that is ≈ 12 further items, for a corrected rate near 25% against the detector’s 30.0%. The flag rate is also uneven along the split: in consecutive blocks of fifty it runs 12%, 44%, 28%, 24% and 44%. The first block is the outlier and the remaining 203 items are flagged at 34.5%, so the unevenness is a step at the start rather than a gradient. We can say no more than that from the release, which carries no provenance column: in particular the step does not fall at item 94, the size of the JF-ICR set EBISU itself evaluates — the first 94 items are flagged at 28.7%, where EBISU’s exclusion of multi-part questions would predict close to zero, so the open split is not that curated set plus an extension. None of this changes the test-split figure, where all 50 questions are single-sentence and none is flagged.

(a) Multi-part question		open split, item 3	Label
Ask 1	25年度の業績目標の背景を教えてください。 'Please explain the background to the FY25 earnings target.'		
↔	今回は、その反省も踏まえて分かりやすい計画を意識した。 'This time, taking that lesson on board, we aimed for a plan that is easy to understand.'		+1
Ask 2	また、通商政策による影響で環境が変わった場合は、収益の下押しを抑制する手段はあるのか。 'And if the environment changes under trade policy, are there means to contain the downward pressure on earnings?'		
↔	環境が大きく変われば、計画を見直すことになる [...] 結局は同様の対応を迫られると考えている。 'If the environment changes materially we will revise the plan [...] we would end up having to respond the same way.'		0
Labelled jointly, with no rule stated			0
Labelled by its most committal part (ours)		gold: +1	+1
(b) Single-part question		open split, item 34	Label
Ask	今後の日銀の金融政策の見通しに対する見解は。 'What is your view on the outlook for Bank of Japan monetary policy?'		
↔	次回利上げが年末年始にあるかはさほど重要視しておらず [...] 'We do not attach much importance to whether the next rate hike comes around the new year [...]'		hedged
↔	いずれにせよ来年夏頃までには利上げがあるだろうと考えている。 'In any case, we think a hike will come by around next summer.'		committal
Rule stated unconditionally			+1
Rule gated on the multi-part detector, so not stated (ours)		gold: 0	0

Table 4: Why the aggregation rule is stated, and why it is gated. In (a) the two asks are answered at different commitment levels, so the item has no determinate label until the prompt says which part governs: taking the most committal part recovers the gold label, labelling the question jointly does not. In (b) a *single* ask draws one long response that nonetheless contains spans at differing commitment levels, and the same directive stated unconditionally latches onto the most committal span, over-labelling an item the annotator read as hedged. Responses are elided to the decisive spans; English glosses are ours.