

Jetsons at the FinEval 2026: Minimal-Pair Augmentation and Label-Routed Ensembling for Greek and Japanese Financial NLP

Parag Pravin Dakle, SaiKrishna Rallabandi, Preethi Raghavan

Fidelity Investments

{paragpravin.dakle, saikrishna.rallabandi, preethi.raghavan}@fmr.com

Abstract

In this work, we describe our submissions to two FinEval-2026 subtasks - Greek financial NER and Japanese implicit commitment recognition where both tasks contain small datasets at 400 and 253 labeled instances. For Subtask 1, we select ensemble models by *error diversity* rather than individual rank, and combine them via **label routing**: each label is taken from the model that performs best empirically on that label. For Subtask 2 we introduce **ordinal minimal-pair augmentation**, synthetically rewriting only the stance-bearing sentence-final clause of a real investor-relations response for a specific label on the commitment scale and verified by two judges from a different model family. Additionally, we perform **style-matched generation** conditioned on the target register and decode under the organizers' published label distribution via **constrained assignment**. Style matching contributes more than every modeling change combined (+0.17). Our leaderboard scores are **0.8879** entity F1 for subtask 1 and **0.8571** accuracy for subtask 2.

1 Introduction

The Natural Language Processing research community has widely studied named entity and intent recognition across multiple languages and domains (Tjong Kim Sang and De Meulder, 2003; Li et al., 2022), including cross-lingual settings that transfer annotation across hundreds of languages (Pan et al., 2017; Conneau et al., 2020). In the financial domain, researchers have studied annual reports and regulatory filings (Loukas et al., 2022; Salinas Alvarado et al., 2015), public company disclosures (Loughran and McDonald, 2011), and investor communications (Keith and Stent, 2019) for these tasks. The FinEval-2026 shared task includes two subtasks focused on these problems: NER over Greek annual reports and implicit commitment recognition over Japanese investor-relations (IR) Q&A. These two share little beyond scale:

400 and 253 labeled instances. In both, the main improvements came from data-side and decoding-side experiments rather than architectural changes. There are four main components of the submitted system across both tasks:

1. Diversity-driven ensemble construction:

The final ensemble solution members are selected based on their performance on each of the named entities rather than overall ranking on the leaderboard. Each model in the ensemble is assigned a weight equal to its overall F1 on the leaderboard, which is used with the predictions for each token to produce the final token label.

2. Label-routed ensembling:

Contrary to the numeric half of the test set, for the text entity half, routing each label to its empirically best ensemble member works the best. This contrast stems from the ensemble having only two members for the textual entity part of the test set.

3. Ordinal minimal-pair augmentation:

For the second subtask, we extend the dataset using counterfactually-augmented data (Kaushik et al., 2020) from a binary flip to a signed *distance* along an ordinal scale, validated by two cross-family judges (§4.1).

4. Constrained (balanced) decoding:

With a known label distribution over the test set, we use joint decoding under a Hungarian assignment with exact per-label quotas at inference time rather than per-test-sample argmax.

2 Task setup

Subtask 1: GRFinNUM. The public release is 320 training and 80 validation sentences from Athens Stock Exchange annual reports (Peng et al., 2025, 2026a), carrying 1,329 numeric entities over five types (App. A). The motivation behind experimentation method design is that the en-

tity type boundaries are annotation conventions rather than lexical characteristics: a bare integer is XPONIKA, ΠΟΣΟΤΗΤΕΣ or ΑΛΛΑ depending on context, and dates are decomposed into separately tagged components, which is why temporal entities are 56% of the pool. A regular expression reaches 0.985 F1 on XPONIKA but 0.312 on ΠΟΣΟΤΗΤΕΣ: the hard part is typing numbers, not finding them, so convention-learning and surface methods fail on disjoint inputs that the observation §3.3 exploits.

Two characteristics of the given dataset are important for the work presented for this subtask. The given label column is aligned to an undocumented subword tokenization we could not reproduce, so we rebuild character spans from the answer column, recovering 98.6% of entities (App. B). Additionally, the 200-row test set is comprised of two subsets: it splits by ID prefix into 100 gr_num_* and 100 gr_text_* rows, the latter being the same sentences doubly annotated under a disjoint label set (ΠΡΟΣΩΠΙΟ/ΟΡΓΑΝΙΣΜΟΣ/ΤΟΠΟΘΕΣΙΑ, i.e. GRFinNER) with no published Greek training data. Therefore, half of the test-set score rests on a label set with zero in-domain supervision.

Subtask 2: JF-ICR. The public set is 253 Japanese IR question/response pairs (Peng et al., 2026b,a) on a five-point ordinal commitment scale, from +2 (strong commitment) to -2 (strong refusal). The experimentation pipeline is determined based on three properties: (1) The commitment scale is unlearnable at its extremes using the given dataset - -2 has two instances and -1 eleven, against 139 for +1, (2) Response length is confounded with label: commitment-side responses average over 330 characters and -2 responses 84, while the 50 test responses average 59 characters against 289 in training, a register gap that penalizes any model learning any length features (§4.2), and (3) The test set label distribution is published: the 50 test items are balanced at 10 per label, split for scoring into a 15-item public and a 35-item private leaderboard, each itself balanced. We treat this as a decoding constraint (§4.3).

3 Subtask 1: Greek Financial NER

3.1 Span Reconstruction

For each (entity, TYPE) line in the answer column, we search text left-to-right for the entity surface string, skipping character positions al-

ready claimed by an earlier entity, and produce (start, end, TYPE, surface). Because entities can repeat components (01.01 → 01, 01), we need claimed-position tracking; otherwise, a naive search resolves every repeat to the same occurrence. This reconstruction yields 1,310/1,329 entities (98.6%) and the remaining 19 are treated as upstream noise. We trained every model on these spans, and the same process at inference guards against prediction failures seen during inference (App. B).

3.2 Experimentation and Results

We evaluated ten methods before ensembling so that member selection could rest on per-label performance rather than a single ranking (App. F). We draw two observations from these results - (1) *Monolingual beats multilingual at this scale:* GreekBERT (Koutsikakis et al., 2020) reaches 0.959 validation micro-F1 against XLM-R’s (Conneau et al., 2020) 0.784 under identical training, a 17.5-point gap consistent with the tokenizer-quality attributed to the monolingual advantage (App. I), and (2) the LLM is the weakest learned method: Qwen3-8B (Qwen Team, 2025) zero-shot is worse than the regular expression (0.872) and few-shot (0.696) trails the CRF (Lafferty et al., 2001) by 25 points, the difficulty being convention-following rather than language understanding.

3.3 Diversity-driven ensembling

In diversity-driven ensembling, each model in the ensemble predicts a tag weighted by its validation F1; the tag with the largest weight wins, and adjacent tokens sharing a tag are merged into a span. The crucial part of this method is which three models the ensemble uses. We selected for complementary strengths rather than rank on the leaderboard: the regular expression is best or tied on the three types with surface cues and performs poorly on the two that need context (ΠΟΣΟΤΗΤΕΣ 0.312, ΑΛΛΑ 0.419), GreekBERT has the opposite weakness at 0.853 on XPHMATA, and the CRF performs evenly across all five labels. The resulting ensemble reaches a score (0.982) which is above every member, and above the two strongest paired together (0.957). The gain lands exactly where the model performances differ, with XPHMATA rising from 0.853 to 1.000 once the other two can outvote the encoder.

3.4 Label-routed ensembling for the text half

For the `gr_text_*` half no Greek person/organization/location data exists in the given dataset, so we bootstrapped some. We use a GreekBERT fine-tuned on wikiann Greek (Pan et al., 2017) with the balanced splits of Rahimi et al. (2019), reaching 0.913 micro-F1 on WikiANN’s held-out split.

The first system we evaluate is a rule tagger plus Qwen3-8B few-shot that reports an F1 of 0.778 against an approximate reference built by translating our predictions to English and fuzzy-matching them to an English version of the same dataset (App. D). Adding the WikiANN classifier via weighted voting degrades performance, from 0.778 to 0.726. The WikiANN classifier’s person-name patterns transfer across domains, but its organization and location predictions do not transfer to Greek financial reports. It strips legal suffixes (A.E.), misses Latin-script company names, and over-tags generic nouns and dates.

Our ensembling strategy here is routing, not weighting. Instead of one weight vector over the members M , we estimate a routing map $\tau : L \rightarrow M$ over the label set L from per-label validation error, and take each label’s spans from $\tau(l)$ alone. Here τ sends $\Pi\text{PO}\Sigma\Omega\Pi\text{O}$ to the WikiANN encoder and $\text{OPTAN}\Sigma\text{MO}\Sigma/\text{TO}\Pi\text{O}\Theta\text{E}\Sigma\text{IA}$ to the LLM+rules member, lifting the combination to 0.789 with no regression: $\Pi\text{PO}\Sigma\Omega\Pi\text{O}$ $0.853 \rightarrow 0.919$, with $\text{OPTAN}\Sigma\text{MO}\Sigma$ 0.791 and $\text{TO}\Pi\text{O}\Theta\text{E}\Sigma\text{IA}$ 0.701 unchanged. While choosing members that disagree is a classical approach (App. I); our point is narrower. A single overall score says how good a member is on average, not where it is good, and averaging across labels hides exactly the differences that make a combination work. The merge rule therefore belongs at whatever level the errors actually differ - in this case, the label. The numeric half makes the same point: XLM-R ranks 7th of 10 overall, yet is the best model on $\Pi\text{O}\Sigma\text{O}\text{THTE}\Sigma$ (0.945).

4 Subtask 2: Japanese Implicit Commitment Recognition

For this subtask, all methods use the same base model, modernbert-ja-310m (Tsukagoshi et al., 2025; Warner et al., 2025), with weighted cross-entropy, and focus on updating the training data or the classification decoder. “Agreement” below means accuracy on the 50 test items against our

hand-read reference labels (§4.4).

4.1 Ordinal minimal-pair augmentation

Back-translation (Sennrich et al., 2016) and paraphrase augmentation (Wei and Zou, 2019) are label-destroying here: the signal lives in sentence-final auxiliaries, and 実施します (+2), 実施したいと考えています (+1) and 実施する予定はありません (-2) differ only in the construction a paraphraser is trained to normalize away. Given an instance (q, r, l) and a target label l' , we rewrite only the stance-bearing clause of r , holding topic, entities and figures fixed, to produce (q, r', l') . This is Kaushik et al.’s (2020) counterfactually-augmented data with two changes forced by the setting: revisions are model-written, and the target is ordinal, so a revision carries a signed direction and a distance $(l' - l)$ rather than simply the opposite label. Conditioning on both keeps -1 and -2 distinguishable instead of collapsing them into a generic refusal register. The label-changing edit also separates these from contrast sets (Gardner et al., 2020), which are built for evaluation; ours enter training.

Each candidate r' is labeled blind (without sight of l') by two judges from a different model family than the generator, and accepted only on unanimous agreement with l' . We specifically do cross-family verification as same-family verification would partly measure self-preference rather than correctness (Panickssery et al., 2024). The overall acceptance rate is quite low, 19.8% for the refusal side and 16.6% for +2; however, we prefer lower acceptance rather than a large dataset with noisy labels.

We tried augmentation only after direct generation failed. Our first attempt resulted in 670 single-turn Q&A candidates from 55 real IR transcripts, ranked them refusal-first with a signal lexicon, and labeled them with a three-judge ensemble. It reproduced the same training label skew rather than correcting it, yielding ten refusal-side instances in total (App. C).

4.2 Style-matched generation

Until this stage of the experimentation, augmentation fixed the label counts but not the writing style: our dataset until this stage averaged 200–340 characters per response against the test set’s 59, so a model could still key on length. We therefore generated 600 more instances matched to the target’s measured length and structural complexity — 120 per label, mean length 67 characters — across a

grid of sectors and topics, with a few test-set inputs shown as style samples (labels are not published, so none were seen). These passed the same two judge verification 85.1% of the time against the minimal pairs’ 19.8%.

4.3 Training and Balanced decoding as constrained assignment

For fine-tuning experiments, we trust the 253 original annotations more than the manufactured data, so every LLM-labelled or LLM-generated instance counts as 0.6 of a gold one, on top of class weights that boost the rare labels and damp the +1 class. The split is by family rather than by instance: a minimal pair is a near-copy of its parent, so separating the two would leave the held-out set measuring memorization. Training uses 1,633 instances, with 51 held out that are gold only and averaged over three seeds (App. E).

At decoding time, let $S \in \mathbb{R}^{50 \times 5}$ be the model’s log-probability matrix over the 50 test items and 5 labels. We expand each label into 10 slots to form a 50×50 matrix and maximise $\sum_i S[i, \pi(i)]$ over perfect matchings π by the Hungarian algorithm (Kuhn, 1955), enforcing exactly 10 items per label globally rather than per item. Applying the constraint over the full test set is what makes it valid under a scoring split the shared task participants cannot see: the 15/35 partition is itself balanced, so a globally balanced prediction is consistent with both halves. This is label-shift correction in its easy case since the label distribution of the test set was published rather than estimated (Saerens et al., 2002; Lipton et al., 2018), and it gained 3-6 items at every generation.

4.4 Model selection

The automatic held-out split created using the given shared task dataset, ended up ranking candidate systems backwards (App. H). Therefore, we used instead a hand-read reference set: we translated the 50 test items to English and labeled them against the published codebook, under an hour of work, then scored candidates by agreement with that reading. This is selection against the evaluation *inputs*, which are public, not the labels, which are hidden. A parallel lexical baseline, character n -gram TF-IDF over a calibrated linear classifier, reached 0.866 on the public slice, matching our second submission and separating from the main experimentation approach only at style-matched generation (App. G).

Evaluation phase	F1	P	R
Public Phase	0.92	0.93	0.90
Private Phase	0.88	0.88	0.88

Table 1: Subtask 1 entity-level leaderboard scores.

Sub.	System	Proxy	Public (15)
v1	v2 model, balanced	0.72	0.733
v2	v3 model, 3-seed, bal.	0.86	0.866
v3	v4 model, 3-seed, bal.	0.9	1.0

Table 2: Subtask 2 submissions. “Proxy” is agreement with hand-read reference over all 50 test items (§4.4).

5 Results

Our Subtask 1 submission routes the two test halves: numeric entities to the weighted vote of GreekBERT + CRF + regex, retrained on the given data; text entities to the rule tagger plus Qwen3-8B, except ΠΠΟΣΩΠΠΟ, routed to WikiANN-GreekBERT. Subtask 2 is a self-contained 310M classifier - ModernBERT-Ja on the style-matched pool, three-seed ensembled, Hungarian decoding.

The final Subtask 2 figure of 0.8571 is 30 of 35 private items; with 15/15 on the public slice this is $45/50 = \mathbf{0.900}$, exactly what the hand-read proxy estimated. This is the strongest evidence for §4.4: the proxy tracked the public score within 0.015 on the first two submissions, then predicted 0.90 for the third against a public readout of 1.000, and the private release settled the discrepancy in its favor. Subtask 1 moved the other way, $0.920 \rightarrow 0.888$, a difference over a few hundred entities that sits inside the expected noise band (Card et al., 2020). We present additional analysis in App. H.

It is important to note that for Subtask-2 no gold test labels were released during or after the shared task and the organizers released only aggregate accuracy on a 15-item public slice after each submission, and no model was trained on, or selected using, those scores. However, the test set (inputs only) was used in the following two ways. First, the test inputs (with labels withheld) were used as unlabeled style anchors when generating the Stage 3 synthetic data, an instance of unsupervised domain adaptation on the input distribution; generated items that near-duplicated an anchor were discarded. Second, to choose between candidates be-

fore making a submission, one author hand-labeled all 50 test items and measured candidate agreement with those labels. These are not gold labels but an inference from the inputs; they were never used for training, and they influenced only selection decisions — which checkpoint, single seed versus ensemble, and whether to blend model generations. The final submitted configuration (stabilized fine-tuning, Stage 3 data, a three-seed ensemble, and decoding under the organizers’ stated 10-per-class prior) was fixed on principled grounds rather than by proxy score, with the sole proxy-driven decision at that stage being a one-item preference for the v4 ensemble alone over a v3+v4 blend. The reported public accuracies (0.733 → 0.866 → 1.000) are official and were obtained on labels the authors never saw.

6 Related Work

Ensembling. That members should be selected for error diversity rather than individual rank is classical (Dietterich, 2000; Kuncheva and Whitaker, 2003). Our §3.4 result is narrower and, we think, less well covered: *weighting* by aggregate validation score is unsafe when a member’s competence varies sharply across labels, and the aggregate gives no signal about that variance. Label routing is a degenerate mixture-of-experts with a hand-estimated, deterministic gate; the small-data regime is what makes the hand estimate preferable to a learned one.

Counterfactual augmentation. Kaushik et al. (2020) introduced counterfactually-augmented data with crowdworker revisions on binary sentiment; Gardner et al. (2020) build minimal edits as contrast *sets* for evaluation. Our extension is the ordinal one: a revision has a direction and a distance, and generation must be conditioned on both or the intermediate classes collapse. The cross-family blind verification gate is what makes model-written revisions usable where Kaushik et al. (2020) could rely on human authorship.

Prior correction. Balanced decoding is label-shift correction (Saerens et al., 2002; Lipton et al., 2018) in the case where the target prior is published, so estimation reduces to a constrained assignment.

Evaluation methodology. Our App. H argument is an applied instance of Card et al. (2020) on underpowered NLP comparisons: with 50 test items,

most claimed improvements are statistically indistinguishable from noise, which is also the right frame for Subtask 1’s 0.920 → 0.888 movement between phases.

LLMs as annotators and judges. Stage 1b rests on evidence that LLM annotation is competitive with crowdsourcing for classification (Gilardi et al., 2023; Ding et al., 2023) — while noting Ding et al.’s finding that the gap widens unfavourably on span-level tasks like NER, consistent with our own Subtask 1 result that an 8B model trails a CRF by 25 F1. Verifying Stage 2 generations with a *different model family* than the generator is supported by Panickssery et al. (2024), who demonstrate a causal link between an LLM’s ability to recognize its own outputs and its preference for them; same-family verification would have measured self-preference as agreement. General judge biases are catalogued by Zheng et al. (2023). Appendix D’s broken validator is a sharper version of the same caution: the failure was not bias but capacity collapse under multi-entity reasoning load.

Monolingual vs. multilingual encoders. The 17.5-point GreekBERT/XLM-R gap in §3.2 is consistent with Rust et al. (2021), who isolate tokenizer quality — not just pretraining data volume — as a driver of monolingual advantage, and with Wu and Dredze (2020) on mBERT’s weakness in lower-resource languages.

7 Conclusion

We described two shared-task systems in which the gains came from the data and the decoder rather than the model: 0.8879 entity F1 on Greek NER and 0.8571 accuracy on Japanese commitment recognition. In each case the novelty was to act at the aspect where the variation was actually observed, rather than where a summary statistic reported it - errors varied by label, so we routed by label; the signal lived in a sentence-final clause, so we edited that clause; the gap was length rather than class balance, so we closed it.

Limitations

Subtask 2’s development numbers rest on 50–51 items with per-class supports as low as one, so differences of a few points are noise; the 0.72/0.86/0.90 proxy figures are in agreement with our own codebook reading, not official gold — the official figure is the 0.8571 in §5. Roughly

two-thirds of the training pool is LLM-labeled or LLM-generated and the benchmark’s strongest published model reaches only ~ 0.61 accuracy, so silver labels carry real noise; our style-matched data in particular is generated to resemble a test set we had inspected, which is legitimate distribution matching but not independent data. Source transcripts are Japanese mega-cap financials and trading houses (2022–2026). Subtask 1’s text-half development numbers are approximate by construction — no Greek gold exists for that half — and WikiANN is silver-standard and general-domain where a manually annotated Greek alternative (eNER; Bartzioakas et al., 2020) existed and was not identified within the competition window; testing eNER-4 as a drop-in for the WikiANN member is the first experiment we would run next, and 0.789 is a floor on that pipeline rather than a ceiling. The numeric pool is 400 sentences of Athens Stock Exchange annual-report Greek only; no claim transfers to other registers or dialects.

References

- Anthropic. 2026. Claude [opus 4.8] system card. <https://www-cdn.anthropic.com/0f0c97ad20d8005706296bd92aa1c27c6b2f4f61/Claude%20opus%204.8%20System%20Card.pdf>. Accessed: [August 27, 2026].
- Nikos Bartzioakas, Thanassis Mavropoulos, and Constantine Kotropoulos. 2020. Datasets and performance metrics for Greek named entity recognition. In *Proceedings of the 11th Hellenic Conference on Artificial Intelligence (SETN)*, pages 160–167.
- Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451.
- Thomas G. Dietterich. 2000. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15.
- Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 11173–11195.
- Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL): Shared Task*, pages 1–12.
- Dimitra Farmakiotou, Vangelis Karkaletsis, John Koutsias, George Sigletos, Constantine D. Spyropoulos, and Panagiotis Stamatopoulos. 2000. Rule-based named entity recognition for Greek financial texts. In *Proceedings of the Workshop on Computational Lexicography and Multimedia Dictionaries (COMLEX)*, pages 75–78.
- Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, et al. 2020. Evaluating models’ local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.
- Ian D. Gow, David F. Larcker, and Anastasia A. Zakolyukina. 2021. Non-answers during conference calls. *Journal of Accounting Research*, 59(4):1349–1384.
- Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations (ICLR)*.
- Katherine A. Keith and Amanda Stent. 2019. Modeling financial analysts’ decision making via the pragmatics and semantics of earnings calls. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 493–503.
- John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. GREEK-BERT: The Greeks visiting sesame street. In *Proceedings of the 11th Hellenic Conference on Artificial Intelligence (SETN)*, pages 110–117.
- Harold W. Kuhn. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97.
- Ludmila I. Kuncheva and Christopher J. Whitaker. 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207.
- John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields:

- Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, pages 282–289.
- Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.
- Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. 2018. Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 3122–3130.
- Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65.
- Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. 2022. FiNER: Financial numeric entity recognition for XBRL tagging. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4419–4431.
- OpenAI. 2026a. Gpt 5.5 system card. <https://deploymentsafety.openai.com/gpt-5-5/change-log>. Accessed: [August 27, 2026].
- OpenAI. 2026b. Gpt 5.6-sol system card. <https://deploymentsafety.openai.com/gpt-5-6/change-log>. Accessed: [August 27, 2026].
- Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1946–1958.
- Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Xueqing Peng, Triantafillos Papadopoulos, Efstathia Soufleri, Polydoros Giannouris, Ruoyu Xiang, Yan Wang, Lingfei Qian, Jimin Huang, Qianqian Xie, and Sophia Ananiadou. 2025. **Plutus: Benchmarking large language models in low-resource Greek finance**. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30176–30202, Suzhou, China. Association for Computational Linguistics.
- 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**. Preprint, arXiv:2602.01479.
- Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 151–164.
- Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP)*, pages 3118–3135.
- Marco Saerens, Patrice Latinne, and Christine Decaestecker. 2002. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41.
- Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. 2015. Domain adaption of named entity recognition to support credit risk assessment. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 84–90.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 86–96.
- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.
- Hayato Tsukagoshi, Shengzhe Li, Akihiko Fukuchi, and Tomohide Shibata. 2025. ModernBERT-Ja. Hugging Face model collection.
- Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope corpus: Biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(S11):S9.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, et al. 2025. Smarter, better, faster, longer: A modern bidirectional encoder

for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2526–2547.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP (Repl4NLP)*, pages 120–130.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, et al. 2024. FinBen: A holistic financial benchmark for large language models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. PIXIU: A comprehensive benchmark, instruction dataset and large language model for finance. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, et al. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*.

A Dataset statistics

GRFinNUM sentences are long (mean 224 characters) and dense (3.4 entities each). In JF-ICR, commitment-side responses average 333–344 characters and –2 responses 84, against a test set whose 50 responses average 59 characters (median 58, min 42, max 78) versus 289 in training.

Entity Type	Train	Val	Share
XPONIKA (temporal)	609	133	55.8%
ΠΟΣΟΣΤΑ (percentage)	202	48	18.8%
XPHMATA (money)	134	34	12.6%
ΠΟΣΟΤΗΤΕΣ (quantity)	81	27	8.1%
ΑΛΛΑ (other)	48	13	4.6%
Total	1074	255	

Table 3: Subtask 1 (GRFinNUM) entity distribution.

Label	Meaning	Count	Share
+2	strong commitment	29	11.5%
+1	weak/qualified commitment	139	54.9%
+0	neutral/hedged	72	28.5%
-1	weak refusal	11	4.3%
-2	strong refusal	2	0.8%

Table 4: Subtask 2 (JF-ICR) class distribution.

B Span reconstruction and a repetition-loop failure

Reconstruction: for each entity, TYPE line in answer, search text left-to-right for the entity string, skipping character positions already claimed by an earlier entity, and emit (start, end, TYPE, surface). Claimed-position tracking is required, not cosmetic: dates legitimately repeat components (01.01 → 01, 01), and a naive search resolves every repeat to the same occurrence. Yield: 1,310/1,329 (98.6%).

The same bug class reappeared at inference. A local 8B model occasionally degenerated into repetition loops, emitting the same entity, TYPE line dozens of times; without claimed-position tracking the reconstruction turned one entity into nine duplicate predictions. The conclusion is that the same defensive routine is needed on both the annotation-parsing and the generation-parsing paths.

C JF-ICR augmentation pipeline

Label yields:

- Stage 1b: +1 371, 0 87, +2 9, –1 8, –2 2
- Stage 2: –2 196, –1 69
- Stage 2b: +2 111
- Stage 3: 120 per label

Models: **Claude-Opus-4.8** (Anthropic, 2026), **GPT-5.5** (OpenAI, 2026a), **GPT-5.6-Sol** (OpenAI, 2026b). All three are used as independent Stage-1b judges (different families, since English-centric models are documented to over-predict commitment on Japanese IR text, so a same-family “unanimous” can be unanimously wrong); **Claude-Opus-4.8** as Stage 2/2b/3 generator; the two GPT models as blind verifiers — deliberately a different family from the generator so the check is independent.

Stage	Procedure	In	Accuracy	Rate
1	Harvest, refusal-rank	—	670	—
1b	3-judge consensus	668	477	71.4%
2	+1/0 $\rightarrow$ -1/ - 2 pairs	1,338	265	19.8%
2b	$\rightarrow$ +2 pairs	668	111	16.6%
3	Style-matched generation	705	600	85.1%

Table 5: Augmentation stages and acceptance rates.

Every synthetic instance carries parent_id, generator id, edit rationale, and per-judge labels, so any accepted instance can be traced to its parent and its verification record.

D Cross-lingual validation for the Greek text half — and its failure

No Greek gold exists for gr_text_*. Our first scorer asked an LLM to align Greek predictions against the English gold directly, and it scored obvious matches — $\text{Λιβαδειάς} \leftrightarrow \text{Livadia}$, $\text{Βουκουρέστι} \leftrightarrow \text{Bucharest}$ — as misses once a row contained more than a few entities. Replacing it with “LLM translates (mechanical), code matches (deterministic, auditable)” moved the *same predictions* from 0.527 to 0.778. **An LLM-as-judge validator can itself be the broken component**, and the failure looks exactly like a bad model. The general rule we drew: give the LLM the context-free mechanical step and keep the aggregation deterministic and inspectable.

E Training configuration

Subtask 1. GreekBERT nlpaueb/bert-base-greek-uncased-v1, 15 epochs, lr 3e-5, batch 8, MPS, ~ 8 min/run. XLM-R base identical settings. CRF (Lafferty et al., 2001): sklearn-crfsuite, L-BFGS, shape/context features, 2.6 s. LLM baselines use Qwen3-8B (Qwen Team, 2025). WikiANN fine-tune: same GreekBERT hyperparameters on 20k Greek Wikipedia sentences.

Subtask 2. sbintuitions/modernbert-ja-310m, batch $4 \times \text{grad-accum } 2$, 8 epochs (early-stopped ~ 4.6), max length 1024, seeds {42, 1337, 2024}. Class weights $(-2/ - 1/0/ + 1/ + 2) = [1.04, 1.59, 1.24, 0.55, 1.28]$; per-sample weight 0.6 on the 1,431 non-gold instances. Split: 1,633 train / 51 gold-only test, with 13 variants of test parents dropped for leakage safety.

v4/seed42 held-out gold detail — accuracy

0.647, macro-F1 0.424, ordinal MAE 0.471; per-class F1: -2 0.00 ($n=1$), -1 0.40 ($n=2$), 0 0.62 ($n=14$), $+1$ 0.76 ($n=28$), $+2$ 0.33 ($n=6$). Note the supports before reading any of these as signal.

F Full Subtask 1 per-label validation F1

This is the table member selection was made from (§3.3).

The complementarity that makes the ensemble work is visible here: regex is best or tied on the three formatted types and near-useless on the two contextual ones; the encoder is the reverse on XPHMATA; the CRF is the most uniform. Note also that XLM-R is the *best* single model on ΠΟΣΟΤΗΤΕΣ (0.945) despite ranking 7th overall — a second instance of the §3.4 point that aggregate rank is the wrong selection signal.

G The parallel lexical system

An alternate approach tried in parallel, this line treated Subtask 2 as flat supervised text classification rather than as an ordinal problem. Features are character n -gram TF-IDF — character-level rather than word-level because Japanese is unsegmented — over a calibrated linear classifier with balanced class handling, emitting predictions directly in the required submission format. The motivation was to capture short lexical, numeric and formatting cues without depending on brittle prompt-only behaviour, and to have a cheap, fully reproducible system that runs anywhere.

Because Subtask 2 is severely class-imbalanced (§2), this line also ran a synthetic augmentation pass, but a conservative one: train-only, *label-preserving* variants of minority-label instances, generated to reduce the classifier’s bias toward the dominant class. Synthetic instances were treated as auxiliary training signal rather than as gold and were never used for evaluation — the same posture as the non-gold down-weighting in §4.3, and the reason model selection here rested on held-out validation behaviour rather than on synthetic-data

Method	XPHMATA	ΠΟΣΟΣΤΑ	XPONIKA	ΠΟΣΟΤΗΤΕΣ	ΑΛΛΑ
ens-top3-w	1.000	0.989	1.000	0.929	0.818
GreekBERT	0.853	0.989	1.000	0.912	0.762
CRF	0.957	0.989	0.960	0.833	0.818
regex	0.954	0.989	0.985	0.312	0.419
XLM-R	0.914	0.854	0.695	0.945	0.720
Qwen3 5-sh	0.686	0.958	0.588	0.926	0.182

Table 6: Per-label validation F1

Method	Micro-F1	Macro-F1
regex heuristic	0.872	0.732
Qwen3-8B, 0-shot	0.308	0.323
Qwen3-8B, 5-shot	0.696	0.668
XLM-R base (fine-tuned)	0.784	0.826
GreekBERT (GB) (fine-tuned)	0.959	0.903
CRF (handcrafted features)	0.947	0.911
GB+CRF unweighted	0.957	0.906
GB+CRF+regex unweighted	0.980	0.950
GB+CRF+regex weighted	0.982	0.947
GB+CRF+regex		

Table 7: Detailed individual system results for Subtask 1

performance alone.

The contrast with §4.1 is deliberate and worth stating: that line manufactures *label-changing* minimal pairs to make the scale’s endpoints learnable at all, while this one augments minority labels *label-preservingly* to rebalance the marginal. The two address the same imbalance from opposite directions. Running both over a single shared split would be a clean ablation, and is one we did not have time to do.

Development numbers: a train-only validation baseline of 37/64 (57.81%), improved to 39/64 (60.94%) by the strongest validation-tuned variant after tuning class weighting and calibration. Two caveats are load-bearing. First, the 64-item split is not the 51-item leakage-safe grouped split used everywhere else in this paper (§4.3), so these figures are not directly comparable to any other number reported here. Second, the tuning and the measurement used that same split, so 60.94% is a deadline-oriented development signal rather than an unbiased estimate of hidden-test performance.

Generation	Pool	1 seed	3-seed+bal.
v1 refusal har-vest	730	0.46	-
v2 + refusal min. pairs	995	0.66	-
v3 + +2 pairs, stab.	1097	0.67	0.86
v4 + style-matched	1697	0.84	0.9

Table 8: Cumulative data generations for Subtask 2.

On the public leaderboard this line reached 0.866 (13/15) — the same score as our second submission (§5), and two items below the submitted system’s 15/15. This supports narrower claim that the augmentation pipeline did not clearly separate from a lexical model until the style-matched generation.

H Analysis

What contributed. Each row of Table 8 adds one component to the row above: cumulative generations, not single-factor ablations. Minimal-pair augmentation of the refusal end is the largest structural fix (+0.20) and is what makes the scale learnable; extending the operator to +2 was nearly flat (+0.01), that class being scarce but not absent. Style matching contributed more than every modeling change combined (+0.17 from 600 instances added to a pool of 1,097), and seed ensembling plus balanced decoding a further +0.06. We spent three generations on the *label* distribution before measuring the *length* distribution, and the second was worth more.

What did not work. Besides blanket weighted voting (−0.052 F1, §3.4) and refusal-first harvesting (§4.1), our first cross-lingual validator asked an LLM to align Greek predictions against English gold directly, and it scored obvious matches as misses once a row held more than a few entities. Replacing it with “LLM translates (mechanical),

code matches (deterministic, auditable)” moved the *same predictions* from 0.527 to 0.778 (App. D). The validator, not the system, was the broken component — and the failure looked exactly like a bad model.

The held-out split was anti-correlated with the target. v1 beat v2 on the held-out gold split (0.71 vs 0.51) and lost on the actual test items (0.46 vs 0.66); v3 beat v4 on the split and lost again. The reason is structural: a gold-only split inherits the 55% +1 marginal, so it rewards collapsing onto the majority class — the very failure that destroys performance on a uniform target. On our best system the held-out numbers look mediocre (0.647 accuracy, 0.424 macro-F1; App. E) while realized test accuracy is 0.900. Where the target distribution is known to differ from the training one, a split built from training data is not a neutral instrument.

I Extended related work

Greek NER resources. Rule-based Greek NER over financial text has direct precedent in [Farmakiotou et al. \(2000\)](#), whose hand-crafted lexical resources over Greek financial news anticipate our regex baseline — and which, at 0.872 F1 in our leaderboard, remains a competitive ensemble member 25 years later. eNER ([Bartziokas et al., 2020](#)) provides manually annotated Greek NER at two granularities: eNER-4 (PERSON/ORGANIZATION/LOCATION/MISC) and eNER-18, which additionally covers MONEY, PERCENT, DATE, QUANTITY, CARDINAL and ORDINAL. Both map onto the shared task better than WikiANN does — eNER-4 onto the `gr_text_*` half directly, and eNER-18’s numeric types onto the `gr_num_*` label set — and the corpus is gold rather than silver. We used neither.

Financial NLP benchmarks. Plutus-ben ([Peng et al., 2025](#)) sits in the lineage of PIXIU ([Xie et al., 2023](#)) and FinBen ([Xie et al., 2024](#)) from the same group.

Hedging and non-answers. JF-ICR’s 0/+1 boundary is close to speculation and hedge detection as posed by the CoNLL-2010 shared task ([Farkas et al., 2010](#)) over the BioScope annotations ([Vincze et al., 2008](#)), a connection the EBISU framing does not make. Independently, the accounting literature has quantified the same phenomenon: [Gow et al. \(2021\)](#) build a linguistic measure of managerial non-answers in earnings conference calls

and find roughly 11% of analyst questions elicit one — a useful external sanity check on any assumed prior over the refusal end of the JF-ICR scale.