

FinEval Challenge @ EMNLP 2026 - FinNLP Workshop: Multilingual Financial Language Understanding

Yupeng Cao¹, Haohang Li¹, Yuechen Jiang², Yueru He³, Yuyan Wang²,
Ruoyu Xiang⁴, Yangyang Yu¹

¹Stevens Institute of Technology, ²The University of Manchester,

³Columbia University, ⁴The FinAI

ycao33, hli113, yyu44@stevens.edu

yuechen.jiang, yuyan.wang-11 @postgrad.manchester.ac.uk

, yh3507@columbia.edu, xry0408@gmail.com

Abstract

Financial language understanding requires accurate interpretation of numerical disclosures, implicit corporate intent, and evidence distributed across languages and documents. We present FinEval 2026, a shared task at the FinNLP workshop at EMNLP 2026¹, comprising Greek financial named-entity recognition, Japanese implicit communication recognition, and multilingual financial question answering. The registration records contain 35 teams after reconciliation, and 14 teams contribute system description papers. The final leaderboards contain 7, 10, and 15 entries for the three sub-tasks, respectively. We describe the task design, participation, submitted approaches, and final results. Leading scores reach 0.9676 entity F1, 0.9714 accuracy, and 0.3827 ROUGE-1 F1. Systems combine domain rules, language-specific modeling, structured prompting, retrieval, and controlled answer revision. Their results highlight the value of task-specific constraints while exposing limitations arising from small test sets, development-to-test shifts, and reuse of previously released reference material. We identify priorities for more reliable multilingual financial evaluation.

1 Introduction and Background

Financial analysis draws on annual reports, financial statements, news, and investor-relations discussions (Xie et al., 2024; Cao et al., 2024; Li et al., 2025; Cao et al., 2025; Wang et al., 2026a). These sources provide complementary information about reported quantities, future intentions, and the context of corporate developments. Understanding them requires preserving units and reporting periods, distinguishing qualified commitments from refusals, and connecting claims to supporting documents. Errors in these operations can change the interpretation of an otherwise fluent answer.

Financial AI has progressed from specialized classifiers toward language models supporting multiple financial tasks. FinBERT adapts pretrained representations to financial sentiment (Araci, 2019), while BloombergGPT trains a generative model on financial and general text (Wu et al., 2023). Evaluation has similarly broadened: FinQA and TATQA test numerical reasoning over reports, tables, and text (Chen et al., 2021; Zhu et al., 2021); FinBen covers extraction, question answering, forecasting, and decision-making (Xie et al., 2024). FinAuditing analyzes structured financial documentation (Wang et al., 2026b). Shared challenges also examine agent-based applications such as trading (Yu et al., 2025), and professional financial services (Han et al., 2026; Cao et al., 2026).

This expansion makes multilingual evaluation increasingly necessary (Peng et al., 2026a). Financial information reflects local reporting practices and linguistic conventions, whereas many evaluation settings emphasize English. Translation facilitates access but does not establish that models preserve inflected entity spans or recover the pragmatic force of corporate answers. Multilingual understanding also requires combining sources: a filing from the first language and a news article in other languages may contribute different evidence for the same question.

Recent benchmarks make these challenges more concrete. Plutus introduces expert-annotated resources for Greek financial language, including numeric and textual entity recognition (Peng et al., 2025). Ebisu evaluates Japanese financial understanding and introduces implicit communication recognition as a distinct problem (Peng et al., 2026b). MultiFinBen extends evaluation across languages and modalities, with PolyFiQA testing the integration of financial statements and multilingual news (Peng et al., 2026a). Together, these resources motivate an evaluation that spans precise extraction, culturally situated interpretation, and

¹<https://thefin.ai/finnlp-2026.html>

evidence-grounded synthesis.

FinEval 2026 turns these capabilities into three shared subtasks, supporting learned encoders, deterministic rules, prompted language models, and multi-stage workflows. This report documents the evaluation setting and participation, summarizes all 14 system papers, and analyzes the final private leaderboards. We distinguish official scores from team-specific experiments, whose development subsets, external resources, and protocols differ.

2 Challenge Overview

Table 1 summarizes the released data and evaluation targets. The subtasks address different levels of financial understanding and have separate rankings without an aggregate score. Task descriptions, data links, rules, and final standings are available on the official website.²

2.1 Greek Financial Named-Entity Recognition

The first subtask draws on Plutus (Peng et al., 2025). GRFinNUM identifies monetary amounts, percentages, temporal expressions, quantities, and other numeric entities; GRFinNER identifies persons, organizations, and locations. Predictions must preserve the source expression and assign its category. Exact entity matching makes boundary conventions consequential, including the treatment of multi-word names and date components.

The public numeric-NER release contains 320 training and 80 validation examples. The test file contains 100 numeric and 100 textual prompts over the same 100 passages: 200 task records but only 100 distinct passages. The absence of corresponding Greek textual-NER supervision motivates different strategies for the two components. Final evaluation uses entity F1, with precision and recall also reported. Numeric-only development scores therefore measure a narrower problem than the combined ranking.

2.2 Japanese Implicit Communication Recognition

The second subtask is Japanese Financial Implicit Communication Recognition (JF-ICR), based on Ebisu (Peng et al., 2026b). Given a financial question and a company response, a system predicts one of five ordered labels: strong commitment

(+2), qualified commitment (+1), neutral or non-directional response (0), qualified refusal (−1), and strong refusal (−2). The target is communicative stance toward the relevant proposition, rather than positive or negative sentiment. Negation, conditions, finality, and the distinction between declining an action and declining to disclose information all affect the label.

The released labeled file contains 253 examples: 29 with +2, 139 with +1, 72 with 0, 11 with −1, and 2 with −2. These counts were checked directly against the public release. Always predicting +1 consequently yields 54.9% accuracy on this development pool. The hidden-label participant file contains 50 examples. System papers consistently describe the announced test design as balanced, with 10 examples per label and scoring partitions of 15 public and 35 private instances (Kapoor et al., 2026; Morita, 2026; Dakle et al., 2026). This is a substantial change from the development distribution. The original benchmark’s 94-example evaluation set, mentioned on the task website, should not be confused with the larger 253-example shared-task release.

Accuracy is the ranking metric, assigning the same penalty to all incorrect labels despite the ordinal scale. Teams additionally report developmental macro-F1, macro-recall, or ordinal error to diagnose class imbalance and boundary errors.

2.3 Multilingual Financial Question Answering

The third subtask uses PolyFiQA from MultiFinBen (Peng et al., 2026a). Inputs combine SEC financial statements with news in English, Chinese, Japanese, Spanish, and Greek. Systems must identify the information requested, reconcile the sources, and produce a concise answer with supporting evidence. The released development and test files each contain 76 questions, organized into 19 financial-report groups with four questions per group. Repeated report identifiers denote shared source material, whereas the test file’s added row identifier is the unique submission key.

Inspection of the released prompts establishes a meaningful change in the required source roles. Development questions ask for an answer from financial statements and corroborating news evidence. Test questions ask for a synthesis of news, supported by quotations from financial statements. Test question families concern revenue focuses, margin strategies, capital allocation, and capital

²<https://www.finnlp2026sharedtask.info/>

Subtask	Languages	Labeled release	Participant test set and output	Ranking metric
Greek financial NER	Greek	320 train + 80 validation	200 prompts over 100 passages; numeric or textual entity/type pairs	Entity F1
Japanese JF-ICR	Japanese	253 examples	50 question–response pairs; one of five commitment/refusal labels	Accuracy
Multilingual PolyFiQA	EN, ZH, JA, ES, EL	76 examples	76 questions in 19 report groups; answer plus financial-statement evidence	Macro ROUGE-1 F1

Table 1: Released shared-task data and evaluation targets. Greek labeled examples cover numeric NER; textual NER has no corresponding labeled Greek training release. Test-set counts include both public and private scoring portions. EN, ZH, JA, ES, and EL denote English, Chinese, Japanese, Spanish, and Greek.

expenditure. The required fields are Answer and Financial Statements Evidence, with a total response limit of 100 words. When an answer or supporting evidence is unavailable, the relevant content should be None. Systems therefore need to adapt both their reasoning and their output conventions across the two releases.

The ranking metric is macro-averaged ROUGE-1 F1 (Lin, 2004), with precision and recall reported separately. The official scoring partitions contain 24 public and 52 private questions, as documented by participating teams (Marti and AlNuaimi, 2026; Atiq and Butt, 2026; Prasanna et al., 2026). Because F1 is averaged at the example level, its reported value need not equal the harmonic mean of aggregate precision and recall. ROUGE captures lexical agreement with the reference answer; it does not independently establish numerical correctness or evidence support. Team-specific grounding checks and judge scores are therefore discussed as supplementary analyses.

Participation and evaluation protocol. Practice data were scheduled for release on July 28, 2026, with the test phase running August 8–24 and system papers due August 27. We use the organizer-verified private standings updated September 3. The rules allow training, fine-tuning, augmentation, prompt rewriting, and external data. They specify a 70-billion-total-parameter limit and OpenRouter for API workflows; mixture-of-experts models are counted by total parameters. We reproduce the published standings without independently adjudicating participant-reported exceptions or implementation details.

3 Participants Overview

The registration spreadsheet contains 49 response records and 37 distinct submitted team-name strings. We consolidate repeated member registrations, merge the punctuation variants

curiosity_exp and curiosity-exp belonging to the same registrant, and reconcile GM with ADML using the registered members and the ADML paper. This gives **35 registered teams**. Task interest is combined across members of each team: 14 teams register for Greek NER, 19 for Japanese JF-ICR, and 31 for PolyFiQA. Twelve teams register for all three subtasks, five for two, and eighteen for one. These are registrations of interest, not proof of a scored submission.

Using the self-declared affiliation-type field, the team-level distribution is **12 industry, 10 academic, and 13 individual teams**. Two consolidated teams contain both Industry and Individual responses; we count these under Industry once, avoiding double counting. Other teams retain their declared participation category even when an individual lists a university or employer. Thus, these figures describe how teams registered rather than an inferred classification of their members’ employment. Figure 1 presents the registration categories and leaderboard participation by subtask.

The three final leaderboards contain **32 task-level entries**: 7 for Greek NER, 10 for JF-ICR, and 15 for PolyFiQA. Their union contains 20 submission identities after normalizing the spelling variation Fin-Agent/Fin_Agent. Several further account-to-team links can be established from paper authors and matching score tuples: gmarti corresponds to ADML, israelkingz to PIO, mabdelmagid to Al-Qahera, and yarry to Finaira. Multiple accounts associated with one team require additional reconciliation before the union can be treated as an exact team count. We retain the displayed identifiers in the result tables so that every score can be checked directly against the official standings.³

Finally, **14 teams contribute 14 system papers** (Table 2). Four describe Greek systems, eight Japanese systems, and nine QA systems; these

³The available records establish scored participation, not the total number of final code-form submissions.

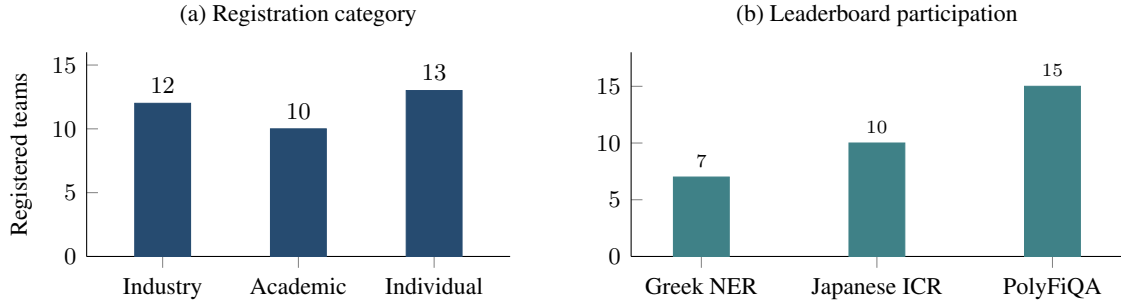


Figure 1: Participation from registration records and final private leaderboards. Categories in (a) are mutually exclusive under the stated reconciliation rule. Panel (b) shows only the final leaderboard entries for each subtask; accounts can differ from team names, so these are not a deduplicated team count.

counts overlap. Paper coverage can be narrower than scored participation: ADML describes QA, shamikkundu describes JF-ICR, and Jetsons describes Greek NER and JF-ICR. We do not extrapolate their documented methods to other tasks.

4 Evaluation Results

We report final private scores and analyze the documented systems. Tables preserve the website’s order and ranks; identical scores remain performance ties despite consecutive positions. Rank differences do not establish statistical significance, and internal ablations across different datasets are not directly comparable.

4.1 Greek Financial NER

Results. The leading system, gmarti, achieves 0.9676 F1, followed by mojo28 at 0.9252 and DDUN at 0.9156 (Table 3). PIO, Al-Qahera, Fin_Agent, and Jetsons contribute Greek system descriptions. The top account is associated with ADML, whose paper covers QA. Because the paper collection provides no Greek-method description for the top three entries, their design choices cannot be inferred from their scores.

Rules and source-constrained extraction. PIO combines Llama 3.3 70B prompting with separate numeric and named-entity processing (Odeajo, 2026). Numeric patterns capture digit groups, temporal expressions, and financial quantities, while the language model copies textual spans from the source. Subsequent rules address category errors, compound locations, legal suffixes, and selected nationality expressions. The paper reports a public-score progression from approximately 0.949 to 0.9748 after post-corrections, followed by 0.9051 privately. This illustrates the usefulness

of annotation-aware corrections while showing that their public improvement does not directly estimate performance on the private portion.

Fin_Agent also separates numeric from textual extraction (Rashid et al., 2026). Ordered rules identify numeric candidates, while high-recall textual candidates are classified by Qwen2.5-32B using passage context. Constraining classification to proposed source spans reduces unsupported entity generation, but makes recall dependent on candidate coverage. Its numeric-only validation rules reach 0.9862 F1, compared with 0.9287 for the reported XLM-R-Large baseline. The combined private score of 0.8884 reflects a broader evaluation, including textual entities for which equivalent labeled training data were unavailable.

Prompt optimization and complementary models. Al-Qahera uses GEPA to optimize instructions for a frozen Gemma task model (Mohamed and Ajlan, 2026). Starting from the organizer’s numeric-NER instruction, the optimized prompt reaches 0.978 on a separately held-out 80-example numeric subset. A hand-improved seed leaves less room for optimization. The submitted system applies the optimized instruction to numeric prompts and the organizer’s zero-shot instruction to textual prompts, obtaining 0.9038 on the combined private evaluation. Its exploratory program-search experiments also show that the best search-set candidate can perform worse on held-out examples. These internal scores should not be presented as official task-wide improvements.

Jetsons approaches the supervision imbalance through learned and deterministic models (Dakle et al., 2026). For numeric entities, it combines GreekBERT, a conditional random field, and regular expressions chosen for complemen-

Paper team	Coverage	Main documented approach
PIO	G, J, Q	Task-specific prompting, source-preserving extraction, deterministic corrections, and constrained QA output.
Maddies NLP	J	Label-blind, proposition-aware evidence extraction followed by ordinal adjudication.
ADML	Q	Exact-query historical-reference retrieval; a separately developed grounded fact compiler and sparse edits.
AhsanBLAH	Q	Category-calibrated few-shot generation with a conservative formatting and citation guard.
Finaira	Q	Dense multilingual retrieval, structured-data extraction, and validated prompt constraints.
shamikkundu	J	Annotation-guideline prompting and schema-constrained confidence elicitation with a 14B model.
Mircea Andrei Razvan	Q	Unified answer/evidence generation, template-aware prompting, and output normalization.
Daikoku	J	Verbalizer scoring, class-count auditing, and forced-choice adjudication.
LangKG	Q	Explicit answering policy with prompt-variant candidates and deterministic reranking.
Jetsons	G, J	Complementary NER ensembles; Japanese minimal-pair augmentation and quota-constrained decoding.
Fin_Agent	G, J, Q	Numeric rules and candidate classification; Japanese cue scoring; retrieval-guided QA.
Token Crit	J, Q	Validation-gated label revisions and source-linked evidence admission.
Al-Qahera	G, J	Reflective prompt optimization around a frozen task model; zero-shot textual NER.
HX-Fin	Q	Precision-first answer refinement and conservative selection among evidence candidates.

Table 2: Coverage and methods in the 14 contributed system papers. G: Greek NER; J: Japanese JF-ICR; Q: PolyFiQA. Coverage indicates what the paper describes, which can be narrower than the team’s scored participation. Each paper is discussed and cited in Section 4.

Rank	Identifier	F1	Prec.	Rec.
1	gmarti	.9676	.9676	.9676
2	mojo28	.9252	.9199	.9306
3	DDUN	.9156	.9145	.9167
4	israelkingz	.9051	.9051	.9051
5	mabdelmagid	.9038	.9167	.8912
6	Fin_Agent	.8884	.8925	.8843
7	Jetsons	.8879	.8868	.8889

Table 3: Final private Greek NER leaderboard. Identifiers and ranks follow the official website.

tary category-level errors. For textual entities, a WikiANN-trained Greek encoder supplies person predictions, while an LLM-plus-rules component supplies organizations and locations. This label-dependent routing avoids assuming that the strongest average model is best for every entity type. The paper reports 0.982 numeric validation F1 and 0.8879 combined private F1. Its textual development analysis uses an approximate reference derived through English alignments, which limits direct comparison with gold-standard evaluation.

Analysis. Across the documented systems, financial NER requires learning annotation conventions alongside entity recognition. Numeric expressions are often easy to locate but difficult to type; textual entities require handling inflection, abbreviations, and corporate names with less supervision. Future per-category and per-component evaluation could distinguish boundary errors, type confusions, and missed candidates that the combined scores cannot

Rank	Identifier	Accuracy	Correct / 35
1	gmarti	.9714	34
1	Fin_Agent	.9714	34
1	ankexu	.9714	34
4	Maddies NLP	.9429	33
5	shamikkundu	.9429	33
6	daikoku	.9429	33
7	israelkingz	.9143	32
8	mojo28	.9143	32
9	Jetsons	.8571	30
10	mabdelmagid	.6571	23

Table 4: Final private JF-ICR leaderboard. Correct counts are implied by the documented 35-item private split and rounded published accuracies. Consecutive website positions at the same accuracy are not performance differences.

resolve.

4.2 Japanese Implicit Communication Recognition

Results. Three entries share the highest private accuracy, 0.9714: gmarti, Fin_Agent, and ankexu (Table 4). The latter corresponds to Token Crit. Maddies NLP, shamikkundu, and Daikoku achieve 0.9429. On the 35-item private portion, the difference between these groups is one correct prediction, or approximately 2.86 percentage points. As with Greek NER, the ADML paper does not describe gmarti’s Japanese method, so the following comparison concerns documented approaches rather than a complete explanation of the top-ranked systems.

Linguistic rules and constrained revision.

Fin_Agent uses an interpretable ordinal scorer with cues for explicit decisions, volition, conditionality, deferral, and refusal (Rashid et al., 2026). Contextual treatment of negation prevents reaffirmation of an existing policy from automatically becoming a refusal. The system includes Sinkhorn normalization under a uniform class prior, but the paper reports that it changes no test predictions. Its 15/15 public and 34/35 private results yield 49/50 overall. The 98.0% overall score in the paper’s abstract is therefore distinct from the 97.14% private ranking score.

Token Crit decomposes decisions into rubric-based and proposition-based candidates, then admits only supported label changes (Xu and Cao, 2026). Separate checks address direction, withholding, commitment strength, and adjacent-label boundaries. On the 253-example development set, its final workflow reaches 62.1% accuracy versus 51.0% for its stronger direct-prompting baseline. The paper’s final boundary audit proposes 75 changes, of which 19 pass the gates; applying all proposals would lose two correct predictions, whereas gated changes gain eleven. This provides a specific rationale for constraining revision, although the component comparison is developmental rather than a private-test ablation.

Evidence decomposition and confidence elicitation.

Maddies NLP first extracts proposition-aligned Japanese evidence without assigning a label, then asks the same Gemma model to adjudicate the original response using that evidence (Kapoor et al., 2026). The intermediate representation preserves negation scope, conditions, actor control, and institutional anchors. Development accuracy increases from 56.92% to 60.08%, but macro-F1 decreases from 0.354 to 0.331 and the paired accuracy improvement is not statistically conclusive in the team’s analysis. Its private score is 0.9429. This combination cautions against treating a small development accuracy gain as uniform improvement across the five labels.

The shamikkundu system instead uses a single 14B model call with the published annotation guidelines and a required confidence value for each label (Kundu, 2026). It selects the highest-scoring label without fitting task parameters. The method reaches the same private accuracy of 0.9429, demonstrating that a compact workflow can be competitive in this setting. Its reported con-

fidence values are elicited model judgments, not independently verified probabilities. The paper also documents development-data quality concerns and distinguishes manual test pseudo-labels from official hidden labels; we preserve that distinction when interpreting its diagnostic experiments.

Known priors, augmentation, and prompt search.

Daikoku combines independent verbalizer scoring with a class-count audit and forced-choice reassessment (Morita, 2026). The announced test balance identifies overrepresented predicted classes, but matching the quota is not sufficient to guarantee correctness: labels can be exchanged between examples without changing class counts. The submitted system attains the required balance and 33/35 private accuracy. Its analysis is useful precisely because it separates a consistency constraint from evidence that individual labels are correct.

Jetsons fine-tunes a Japanese encoder using ordinal minimal pairs and synthetic examples matched to the test register (Dakle et al., 2026). Label-changing edits target the stance-bearing clause, and two judges from another model family verify the intended label. Joint decoding then enforces the published class quotas through assignment. The system scores 0.8571 privately. Its reported style-matching gains rely partly on the authors’ manual reference labels, so they are evidence about the team’s development process rather than organizer-certified test ablations.

PIO uses rubric-aligned few-shot prompting and targeted lexical corrections, including cases where declining a proposed change reaffirms an existing policy (Odeajo, 2026). It achieves 0.9143 privately. Al-Qahera uses reflective prompt optimization to sharpen distinctions between hedged commitment, neutrality, and indirect refusal (Mohamed and Ajlan, 2026). Its 0.692 accuracy on a team-held-out 52-example development partition is separate from its official private score of 0.6571.

Analysis. The methods converge on proposition scope and commitment strength through different mechanisms. Their high accuracy concerns a small, balanced test set whose responses are reported to be shorter or simpler than development responses; several systems also exploit the published class prior. Paired significance and private class-specific errors cannot be established without item-level predictions and labels. Future evaluations should include larger naturally distributed samples, balanced diag-

Rank	Identifier	R-1 F1	Prec.	Rec.
1	hx	.3827	.3902	.4057
2	gmarti	.3726	.5698	.3542
3	AhsanBLAH	.3702	.3111	.5000
4	DAIR	.3665	.3431	.4224
5	Mircea Andrei Razvan	.3653	.3517	.4194
6	yarry	.3593	.3683	.3771
7	mojo28	.3590	.3392	.4029
8	shamikkundu	.3478	.3122	.4191
9	Fin-Agent	.3419	.3227	.3899
10	dm015	.3282	.2872	.4180
11	LangKG	.3240	.3042	.3897
12	WB	.3062	.3971	.2709
13	Mohotarema	.3045	.3535	.3096
14	Jetsons	.2758	.3106	.2672
15	israelkingz	.2036	.3148	.1864

Table 5: Final private PolyFiQA leaderboard. R-1 denotes ROUGE-1. Scores are macro-averaged by example. The table preserves all displayed account entries, including possible within-team duplicates.

nostic subsets, and explicit disclosure of available batch-level priors.

4.3 Multilingual Financial Question Answering

Results. HX-Fin, listed as hx, leads with 0.3827 ROUGE-1 F1, followed by ADML/gmarti at 0.3726 and AhsanBLAH at 0.3702 (Table 5). The first-to-second gap is 0.0101; second and third differ by 0.0024. These are descriptive differences, not established significant improvements. QA metrics also have a different interpretation from the extraction and classification metrics, so their lower numerical values cannot be read as a cross-task ranking of difficulty. Nine papers describe QA methods, offering evidence about alternative ways to select, express, and verify financial information.

Conservative answer and evidence refinement. HX-Fin separates answer construction from evidence selection (Li and Wang, 2026). It creates candidates from full-context, news-only, reasoning-enabled, multilingual, and retrieval-based views. An order-balanced comparison selects a conservative core; supported additions from a broader candidate pool then pass omission checks and claim pruning. A separate module produces and refines statement evidence, retaining the default unless an alternative is judged more reliable. The paper reports stage-wise improvements under an LLM judge on a 0–4 scale. These diagnostics explain the intended behavior of the refinement stages, but they are not ROUGE-1 ablations and do not independently attribute the winning private score to each

component.

AhsanBLAH uses category-calibrated few-shot generation followed by a constrained formatting and citation guard (Atiq and Butt, 2026). The guard removes extraneous material and checks quoted evidence while preserving the generated content. The paper reports 0.4819 development ROUGE-1 F1 and 0.3702 privately. Because the same 76 development examples support prompt calibration and evaluation, the former is an in-sample result. The team does not provide a controlled guard-versus-no-guard ROUGE ablation, so the guard’s intended reliability benefit should not be described as a separately measured accuracy gain.

Retrieval and output control. Finaira compares sparse, dense, hybrid, and no-retrieval configurations, then combines dense retrieval with structured extraction and constrained generation (Nasser and Mohamed, 2026). Its analysis identifies English-only citation bias and rejects a category-conditioned router that fails cross-validation. The final private score is 0.3593. The development and deployed configurations differ, however, so the paper’s 0.412 cross-validated development score is not an out-of-sample estimate for the exact final system.

Fin_Agent retrieves financial-statement and news chunks with BGE-M3, places this evidence alongside the full context, and generates per-question answers with Qwen3-32B (Rashid et al., 2026). Rewriting enforces concision and English output, while checks flag unsupported figures and quotations. It scores 0.3419 privately. The design illustrates retrieval as an attention aid rather than a replacement for the source context. Its reported experiments also distinguish retrieval, per-question generation, and length normalization instead of attributing all changes to a single component.

Mircea Andrei Razvan uses one generation call for both answer and evidence, with question-template-specific prompts and deterministic normalization (Mircea Andrei Razvan, 2026). The system scores 0.3653 privately. The paper reports that splitting answer and evidence generation produced more missing-evidence outputs, a useful counterpoint to the multi-stage systems. It also discloses that a separate resource used in development later proved to overlap the test set. Consequently, its post-hoc model comparisons against that resource cannot serve as independent held-out evidence.

LangKG formalizes an answering policy cover-

ing source support, exact values, reporting periods, multilingual alignment, and abstention (Prasanna et al., 2026). Its submitted system generates prompt variants and selects a response using deterministic checks for structure, evidence, derivation language, and length. It scores 0.3240 privately. Token Crit instead builds an explicit evidence state and admits records only after checks on attribution, the requested relation, and numerical grounding (Xu and Cao, 2026). Its corresponding private score is 0.3062. On development data, Token Crit substantially improves evidence-section overlap while changing overall ROUGE-1 only from 0.3399 to 0.3425. Better intermediate evidence therefore does not automatically translate into a comparable improvement in the complete answer.

Historical references and evaluation provenance. ADML’s selected submission retrieves exact or conservatively normalized full-query matches from a previously public PolyFiQA snapshot (Marti and AlNuaimi, 2026). The team reports coverage of all 76 released test queries, a public score of 0.9994, and a private score of 0.3726. It separately develops a source-grounded fact compiler with numerical provenance checks and sparse candidate edits. The compiler’s historical results and public score do not establish its private performance, since the reported private result belongs to the retrieval submission. The paper interprets the public/private contrast as evidence of reference-version shift, while acknowledging that hidden answers were unavailable to verify the mechanism directly.

PIO likewise reports that alignment with public Expert references yields a public score of 1.0000, while its described generative private result is 0.2036 (Odeajo, 2026). The paper explicitly distinguishes reference alignment from generative QA competence. These disclosures are relevant under an external-data-permitted competition: they do not by themselves establish a rule violation, but they limit what a high public score demonstrates. In particular, withholding the current answer column is insufficient to establish a novel-input test when related query packets or historical answers have already been released.

Analysis. The QA submissions balance useful evidence coverage against unsupported additions through retrieval, structured generation, and conservative revision. A unified generator remains competitive, while more elaborate evidence processing sometimes yields modest end-to-end gains.

Differences in base models, data access, and tuning prevent the standings from isolating the effect of workflow depth or model size.

Future editions should preserve company/report groups in validation splits, publish versioned data manifests, and audit query and document overlap. Separate checks for answer content, numerical consistency, and evidence support would complement lexical overlap. Multiple references or expert judgments could capture legitimate variation in synthesis. Aggregate scores alone do not establish performance differences by evidence language.

5 Conclusion

FinEval 2026 evaluates three complementary capabilities in multilingual financial language understanding: Greek entity extraction, Japanese implicit intent recognition, and financial QA using multilingual evidence. The shared task attracts 35 registered teams and 14 system papers, with final leaderboards containing 7, 10, and 15 entries. The documented approaches show how domain rules, language-specific representations, structured prompting, retrieval, and controlled revision can be combined to address distinct financial phenomena.

The results also clarify the limits of the current evaluation. Strong scores on a small balanced classification set do not establish robustness to naturally distributed corporate communication, and historical-reference access complicates the interpretation of QA overlap scores. Future work should expand private evaluation, separate numeric and textual extraction results, strengthen data-version and overlap controls, and assess factual support alongside answer similarity. These improvements would help distinguish adherence to a benchmark’s conventions from transferable multilingual financial understanding.

Limitations

Our analysis relies on aggregate leaderboard results and participant reports, without access to private instance-level labels or independent reruns of every system. Small evaluation sets, differences between development and test distributions, and team-specific tuning protocols limit statistical and cross-system comparisons. ROUGE-1 measures lexical overlap and does not fully capture financial correctness, evidence quality, or performance by language.

References

- Dogu Araci. 2019. [FinBERT: Financial sentiment analysis with pre-trained language models](#). arXiv:1908.10063.
- Muhammad Ahsan Atiq and Mohib Ahmad Butt. 2026. [AhsanBLAH at the FinEval 2026: Grounded Two-Stage Financial QA with LLM-Judge Verification](#). In *Proceedings of the 11th Workshop on Financial Technology and Natural Language Processing*, pages 745–751. Association for Computational Linguistics.
- Yupeng Cao, Zhi Chen, Qingyun Pei, Nathan Lee, KP Subbalakshmi, and Papa Momar Ndiaye. 2024. Ecc analyzer: Extracting trading signal from earnings conference calls using large language model for stock volatility prediction. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 257–265.
- Yupeng Cao, Haohang Li, Weijin Liu, Wenbo Cao, Anke Xu, Lingfei Qian, Xueqing Peng, Minxue Tang, Zhiyuan Yao, Jimin Huang, et al. 2026. Fintrace: Holistic trajectory-level evaluation of llm tool calling for long-horizon financial tasks. *arXiv preprint arXiv:2604.10015*.
- Yupeng Cao, Haohang Li, Yangyang Yu, Shashidhar Reddy Javaji, Yueru He, Jimin Huang, Qianqian Xie, Fabrizio Dimino, Xiao-yang Liu, KP Subbalakshmi, et al. 2025. Finaudio: A benchmark for audio large language models in financial applications. *arXiv preprint arXiv:2503.20990*.
- Zhiyu Chen, Wenhu 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Parag Pravin Dakle, SaiKrishna Rallabandi, and Preethi Raghavan. 2026. [Jetsons at the FinEval 2026: Minimal-Pair Augmentation and Label-Routed Ensembling for Greek and Japanese Financial NLP](#). In *Proceedings of the 11th Workshop on Financial Technology and Natural Language Processing*, pages 790–800. Association for Computational Linguistics.
- Yi Han, Yan Wang, Lingfei Qian, Haohang Li, Yupeng Cao, Yueru He, Xueqing Peng, Nanhan Shen, Yitao Xu, Yankai Chen, Dongji Feng, Jimin Huang, Xue Liu, Jian-Yun Nie, and Sophia Ananiadou. 2026. [Can llm agents be cfos? benchmarking long-horizon resource allocation in an uncertain enterprise environment](#). *Preprint*, arXiv:2603.23638.
- Madhav Kapoor, Aditya Channa, and Jastej Singh. 2026. [Maddies NLP at the FinEval 2026: Proposition-Aware Tagged Reasoning for Japanese Implicit Communication Recognition](#). In *Proceedings of the 11th Workshop on Financial Technology and Natural Language Processing*, pages 735–739. Association for Computational Linguistics.
- Shamik Kundu. 2026. [shamikkundu at the FinEval 2026: Zero-Parameter Confidence Elicitation for Japanese Implicit Communication Recognition](#). In *Proceedings of the 11th Workshop on Financial Technology and Natural Language Processing*, pages 760–767. Association for Computational Linguistics.
- Haohang Li, Yupeng Cao, Yangyang Yu, Shashidhar Reddy Javaji, Zhiyang Deng, Yueru He, Yuechen Jiang, Zining Zhu, KP Subbalakshmi, Jimin Huang, et al. 2025. Investorbench: A benchmark for financial decision-making tasks with llm-based agent. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2509–2525.
- Peng Li and Yinchun Wang. 2026. [HX-Fin at FinEval 2026: Precision-Anchored Answer Refinement and Multi-Model Evidence Selection for Multilingual Financial Question Answering](#). In *Proceedings of the 11th Workshop on Financial Technology and Natural Language Processing*, pages 825–830. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Gautier Marti and Khaled AlNuaimi. 2026. [ADML at the FinEval 2026: Version-Aware Retrieval and Grounded Sparse Editing for Multilingual Financial QA](#). In *Proceedings of the 11th Workshop on Financial Technology and Natural Language Processing*, pages 740–744. Association for Computational Linguistics.
- Mircea Andrei Razvan. 2026. [Mircea Andrei Razvan at the FinEval 2026: A Unified Prompting Approach for Multilingual Financial Q&A](#). In *Proceedings of the 11th Workshop on Financial Technology and Natural Language Processing*, pages 768–772. Association for Computational Linguistics.
- Mohamed Mohamed and Hazem Ajlan. 2026. [Al-Qahera at the FinEval 2026: GEPA Prompt Evolution for Greek Financial NER and Japanese Implicit Communication Recognition](#). In *Proceedings of the 11th Workshop on Financial Technology and Natural Language Processing*, pages 819–824. Association for Computational Linguistics.
- Kyosuke Morita. 2026. [Daikoku at FinEval 2026: Multi-Stage Labelling with Forced-Choice Adjudication](#). In *Proceedings of the 11th Workshop on Financial Technology and Natural Language Processing*, pages 773–782. Association for Computational Linguistics.
- Yara Nasser and Ashraf Mohamed. 2026. [Finaira at the FinEval 2026: Dense Retrieval and Validated Prompt Engineering for Cross-Lingual Financial Question Answering](#). In *Proceedings of the 11th Workshop on*

- Financial Technology and Natural Language Processing*, pages 752–759. Association for Computational Linguistics.
- Israel Olanrewaju Odeajo. 2026. [PIO at the FinEval 2026: An OpenRouter API Pipeline for Multilingual Financial Understanding](#). In *Proceedings of the 11th Workshop on Financial Technology and Natural Language Processing*, pages 730–734. Association for Computational Linguistics.
- 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, Haoqiang 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:2602.01479.
- Shivika Prasanna, Jaeseong Lee, Hui Su, and Yourong Xu. 2026. [LangKG at the FinEval 2026: Answer-Contract Knowledge over Instruction Engineering](#). In *Proceedings of the 11th Workshop on Financial Technology and Natural Language Processing*, pages 783–789. Association for Computational Linguistics.
- Mohotarema Rashid, Tanzina Hossain, Xiaoying Song, and Lingzi Hong. 2026. [Fin_Agent at FinEval 2026: Task-Aware Methods for Multilingual Financial NLP](#). In *Proceedings of the 11th Workshop on Financial Technology and Natural Language Processing*, pages 801–811. Association for Computational Linguistics.
- Yan Wang, Lingfei Qian, Xueqing Peng, Yang Ren, Keyi Wang, Yi Han, Dongji Feng, Fengran Mo, Shengyuan Lin, Qinchuan Zhang, Kaiwen He, Chenri Luo, Jianxing Chen, Junwei Wu, Chen Xu, Ziyang Xu, Jimin Huang, Guojun Xiong, Xiao-Yang Liu, and 2 others. 2026a. [Fintagging: Benchmarking llms for extracting and structuring financial information](#). Preprint, arXiv:2505.20650.
- Yan Wang, Keyi Wang, Shanshan Yang, Jaisal Patel, Jeff Zhao, Fengran Mo, Xueqing Peng, Lingfei Qian, Yankai Chen, Víctor Gutiérrez-Basulto, et al. 2026b. [Finauditing: A financial taxonomy-structured multi-document benchmark for evaluating llms](#). In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3456–3463.
- Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kam-badur, David Rosenberg, and Gideon Mann. 2023. [BloombergGPT: A large language model for finance](#). arXiv: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, and 15 others. 2024. [FinBen: A holistic financial benchmark for large language models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 95716–95743. Curran Associates, Inc.
- Anke Xu and Wenbo Cao. 2026. [Token Crit at the FinEval 2026: Role-Specialized, Validation-Gated LLM Workflows for JF-ICR and Multilingual Financial Q&A](#). In *Proceedings of the 11th Workshop on Financial Technology and Natural Language Processing*, pages 812–818. Association for Computational Linguistics.
- Yangyang Yu, Haoqiang Li, Yupeng Cao, Keyi Wang, Zhiyang Deng, Zhiyuan Yao, Yuechen Jiang, Dong Li, Ruey-Ling Weng, and Jordan W. Suchow. 2025. [FinNLP-FNP-LLMFinLegal @ COLING 2025 shared task: Agent-based single cryptocurrency trading challenge](#). In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 401–406, Abu Dhabi, UAE. Association for Computational Linguistics.
- 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, Online. Association for Computational Linguistics.