# Financial Entity Extraction as Selective Language Processing

Judith Watson, Olivia Brooks, Frances Kelly*

* Corresponding author: Frances Kelly

Review Article | Enterprise, Policy and Economic Dynamics | IASCI Research Publishing | Published 23 September 2026

## ABSTRACT

This evidence synthesis examines domain-shift abstention in filings, news, and social media. It argues that the appropriate object of evaluation is the trajectory from tokenization and draft generation through scoring, revision, compression, and release, not a model score in isolation. The review brings together the assigned studies with established work on uncertainty, robustness, provenance, and governance. Across these literatures, a common problem emerges: speed and fluency can conceal semantic loss, correlated self-evaluation errors, or domain-specific failure. The proposed framework separates evidence quality, model behavior, decision policy, and operational monitoring, then asks how each layer changes under distribution shift, adversarial pressure, or incomplete information. It recommends evaluation by slices and repeated trials, explicit reject and escalation policies, preservation of data and reasoning lineage, and prospective monitoring tied to defined actions. The result is a research agenda for systems that are efficient enough to use but also bounded enough to audit. No new experiment is claimed; the article develops a comparative conceptual model and identifies tests that would make future empirical claims more credible.

## KEYWORDS

financial entity extraction; selective language processing; evaluation; monitoring; enough; financial; entity

## Introduction

Which confidence signals remain meaningful when textual register changes? Answering the question requires attention to the full pipeline, not just the predictive model. Its behavior reflects a chain of design decisions about evidence, representation, alternatives, resource use, and response. In filings, news, and social media, these choices interact with institutions and physical processes that the estimator only partially represents. A high score can coexist with fragile inputs, an evaluator that rewards the wrong surface feature, or an action rule that turns modest uncertainty into a costly decision. This makes it useful to treat domain-shift abstention as an evidence problem. The inquiry focuses on the minimum evidence and comparison set required for a credible claim. In filings, news, and social media, this distinction should be tested before it changes an operational decision.

This review follows the inference process from source to action. Its unit is the trajectory from tokenization and draft generation through scoring, revision, compression, and release. This scope makes it harder to commit a familiar error: attributing every failure to model architecture while ignoring data capture, labeling, retrieval, validation, interface design, or organizational incentives. This framing places the focal studies on comparable evidential terms. Although their application settings differ, they each make some part of an inference chain more explicit - a reward signal, a graph relation, a physical boundary condition, a confidence gate, or a revision trigger. Evaluation must establish whether that explicit component remains meaningful when source texts, alternative drafts, reward signals, confidence estimates, domain corpora, and human judgments change, and whether the proposed safeguards lead to selective revision, distillation, abstention, expert review, and continuous validation rather than merely producing another metric. Seen through domain-shift abstention, the issue is not merely technical; it determines which claims remain open to challenge.

## Conceptual Foundations

The framework next distinguishes ordinary variation from missing knowledge and from actors who adapt to the application. Aleatory variation concerns irreducible differences among cases. Epistemic uncertainty concerns missing knowledge, limited samples, or unsupported regions of the input space. Strategic adaptation occurs when users, adversaries, markets, or institutions respond to the application itself. These sources demand different controls. Repeated measurement can characterize variation; new evidence and conservative extrapolation can reduce epistemic uncertainty; red teaming and feedback-aware testing address adaptation. Combining them into one confidence score hides which intervention is appropriate. A defensible system therefore reports uncertainty in a form tied to an available response in place of presenting confidence as a decorative number. Seen through domain-shift abstention, the issue is not merely technical; it determines which claims remain open to challenge.

Four analytically distinct layers organize the problem. The evidence layer records observations and their provenance. The representation layer turns those observations into features, graphs, tokens, or simulated states. The inference layer produces predictions, explanations, translations, anomaly scores, or candidate actions. The decision layer maps those outputs to selective revision, distillation, abstention, expert review, and continuous testing. Reliability can fail at any boundary. Better inference cannot repair evidence that omitted the relevant condition, and a calibrated probability cannot rescue an action rule whose costs were never specified. Conversely, a conservative decision policy may reduce harm even when the underlying model remains imperfect. This layered view makes domain-shift abstention testable because each claim can be assigned to a component and a measurable transition. For the present focus, domain-shift abstention therefore becomes a criterion for deciding which evidence deserves further scrutiny.

## What the Core Studies Establish

SAINF: Intrinsic Self-Correction for Robust Machine Translation with Large Language Models asks how a language model can identify and repair fragile translations without an external quality-estimation model. It uses a conditional workflow that produces a chain-of-dictionary translation, self-scores it, and triggers targeted term interpretation, validation, and revision below a threshold. The relevant evidence is that experiments on ChallengeWMT across language pairs report stronger quality and robustness than prompting and interpretation baselines. The method makes self-correction selective, limiting expensive revision to cases the model judges uncertain. In the present review, this work matters because domain-shift abstention cannot be evaluated as an abstract virtue; it must change how evidence is collected and how an action is withheld. Self-assessment can share the generator's blind spots, and threshold behavior may shift across languages, domains, and model families. (Zhang et al. 2026).

A useful test case is DRAFT-RL: Multi-Agent Chain-of-Draft Reasoning for Reinforcement Learning-Enhanced LLMs. Rather than treating its output as an isolated score, the study frames how multi-agent reinforcement learning can preserve structural diversity instead of repeatedly refining a single answer. Its central mechanism is agents generate several chain-of-draft trajectories, peer agents and a learned reward model select promising paths, and actor-critic updates feed those selections into later reasoning, and its evidential basis is that the paper evaluates code synthesis, symbolic mathematics, and knowledge-intensive question answering and reports gains in accuracy and convergence. It treats alternative drafts as an explicit exploration space rather than incidental sampling noise. For filings, news, and social media, the transferable lesson is methodological: a system should preserve the path from signal to decision and expose the conditions under which that path may fail. The boundary is equally important. Peer agreement can amplify correlated errors, and additional drafts increase inference and evaluation costs unless diversity is measured rather than assumed. (Li et al. 2026).

Evidence for domain-shift abstention becomes concrete in Reliable Financial Named Entity Recognition under Domain Shift. The paper addresses which confidence signals permit safe abstention when financial named-entity recognition crosses textual domains through BERT and LoRA-tuned compact language models evaluated across filings, financial news, and extreme out-of-domain social media using five confidence signals, multiple seeds, and bootstrap intervals. Its evaluation is informative because span probability and self-consistency remain more robust under shift than whole-output probability, while severe shift defeats useful clean-subset recovery. The study separates upstream shift detection from downstream selective prediction instead of treating confidence as domain invariant. This does not make the method a universal component; it identifies a design hypothesis that must be re-tested in the article's focal setting. In particular, three registers cannot represent the full diversity of financial text, entity schemas, languages, or time-dependent market vocabulary. The appropriate use of the result is therefore bounded, comparative, and explicit about unresolved deployment

conditions (Zheng et al. 2026).

## Relevant Foundations

Several established literatures clarify the design space. Domain-adapted language modeling demonstrates why financial vocabulary cannot be treated as generic prose (Araci 2019). Financial text analysis shows that common-language sentiment resources can misread domain-specific terms (Loughran and McDonald 2011). Continued pretraining can improve domain fit but may also narrow coverage if deployment domains remain heterogeneous (Gururangan et al. 2020). Transformer attention provides the architectural basis for long-range contextual modeling in modern language systems (Vaswani et al. 2017). Together these findings imply that domain-shift abstention should be evaluated against explicit alternatives and failure costs. In Financial Entity Extraction as Selective Language Processing, this literature supplies a specific test of domain-shift abstention within filings, news, and social media.

The evidence base offers complementary tests rather than one universal metric. Subword modeling addresses rare forms while also making tokenization a consequential design choice (Sennrich, Haddow, and Birch 2016). BLEU made corpus-level comparison scalable but cannot represent every semantic or cultural error (Papineni et al. 2002). Learned translation metrics improve correlation with human judgments while introducing model and domain dependencies of their own (Rei et al. 2020). They also caution against interpreting one benchmark, one split, or one explanatory artifact as a complete validation argument. In Financial Entity Extraction as Selective Language Processing, this literature supplies a specific test of domain-shift abstention within filings, news, and social media.

A credible assurance case can be built from findings that operate at different levels. Self-consistency uses diverse reasoning samples as evidence, but agreement can still reflect shared bias (Wang et al. 2023). Human-feedback training demonstrates the power and specification sensitivity of learned reward models (Ouyang et al. 2022). Direct preference optimization simplifies alignment training while retaining dependence on preference-data quality (Rafailov et al. 2023). Their combined value lies in making assumptions visible enough to challenge before deployment and to revise after failure. In Financial Entity Extraction as Selective Language Processing, this literature supplies a specific test of domain-shift abstention within filings, news, and social media.

## A Traceable System Architecture

The evidence architecture for filings, news, and social media begins with typed evidence. Each record should identify its source, time, collection conditions, transformations, and relation to other records. Graphs are useful when relations carry meaning, but graph construction is itself a hypothesis. An edge may denote temporal adjacency, physical causation, code dependency, shared infrastructure, or analyst assertion; treating these as interchangeable creates false certainty. The architecture should therefore maintain edge types and confidence, retain alternative links when evidence is ambiguous, and version both the data schema and the rules that construct the graph. A model may aggregate this structure, but the original evidence must remain recoverable for audit. This point narrows the research task for filings, news, and social media: compare alternatives under the same information and resource constraints.

Resource allocation should respond to ambiguity and consequence. Easy, well-supported cases can take a fast path. Shifted, ambiguous, or high-cost cases should activate additional precision, retrieval, alternative drafts, simulation, or expert review. This conditional design is preferable to applying maximum computation everywhere because resource cost is part of implementation quality. It is also preferable to an unconditional fast path because efficiency can amplify a systematic error at scale. The trigger must be evaluated as carefully as the expensive model: coverage, false reassurance, subgroup effects, latency, and the consequences of abstention all matter. The output is not simply a prediction but a package containing evidence links, uncertainty, the action taken, and the conditions that caused escalation. This point narrows the research task for filings, news, and social media: compare alternatives under the same information and resource constraints.

## Testing Claims Rather Than Scores

Evaluation begins by writing each claim in testable form. Each intended claim - such as improved robustness, safer abstention, faster updating, or better structural localization - needs a target population, comparator, outcome, and boundary condition. Random splits are insufficient when related samples share a patient, repository, instrument run, season, organization, or simulated design family. Grouped and temporal splits better approximate the novelty encountered after operational use. Stress tests should vary one factor at a time when attribution is important, then combine factors to measure interaction. Repeated runs are necessary whenever decoding, optimization, retrieval, or numerical kernels can vary. Reporting only the best seed or a pooled mean makes operational instability disappear. This point narrows the research task for filings, news, and social media: compare alternatives under the same information and resource constraints.

The action and its costs should determine the reported metrics. Discrimination measures whether cases are ranked correctly; calibration asks whether confidence corresponds to observed frequency; selective risk asks what error remains after deferral; robustness measures performance across defined perturbations; and utility assigns costs to actions. These are not interchangeable. For domain-shift abstention, the minimum report should include overall performance, slice-level results, uncertainty intervals, coverage-risk curves, stability across runs, and failure examples linked back to source texts, alternative drafts, reward signals, confidence estimates, domain corpora, and human judgments. If the deployed pipeline updates incrementally, the validation must also test forgetting, stale evidence, and rollback. If an automatic judge supplies labels, a sample needs independent human review and content-preserving perturbations that reveal whether the judge is grading substance. This requirement gives practical content to the article's central question: Which confidence signals remain meaningful when textual register changes?

## Deployment Controls

Evidence must eventually become a set of enforceable operating conditions. A operational use specification should name allowed uses, prohibited uses, escalation thresholds, data-retention rules, and the person or role that can halt the deployed workflow. Monitoring should track input shift, missingness, abstention, disagreement, latency, overrides, and downstream outcomes. A rising override rate may indicate model drift, a changed workflow, or new user behavior; treating it only as operator noncompliance wastes evidence. Incident review should reconstruct the entire chain, including the model and the surrounding interface. Versioned models without versioned prompts, retrieval indexes, rules, and datasets do not support reliable reconstruction. This point narrows the research task for filings, news, and social media: compare alternatives under the same information and resource constraints.

Placing a person in the loop does not by itself create effective control. Reviewers need enough context to challenge the output, but not so much generated rationale that weak evidence is obscured by volume. Escalation queues require prioritization and workload limits. A reject option that sends nearly everything to experts is safe only on paper; a reject option that rarely fires may be equally ineffective. For filings, news, and social media, oversight should therefore be evaluated with realistic staffing, time pressure, and disagreement. The system should record the final action and its reason without turning that record into surveillance unrelated to the stated purpose. Contestability, privacy, and security are properties of this institutional process as much as of the predictive component. This requirement gives practical content to the article's central question: Which confidence signals remain meaningful when textual register changes?

## Experiments the Field Still Needs

Future validation sets should be organized around plausible mechanisms of failure. Cases should represent unsupported regions, ambiguous evidence, conflicting modalities, rare but costly conditions, and realistic adversarial transformations. Each case needs provenance and a reason for inclusion. The second priority is intervention testing: when uncertainty is detected, compare additional computation, retrieval, alternative reasoning paths, model ensembles, and human escalation under the same resource budget. This reveals whether a proposed safeguard actually changes the decision or only changes the explanation. The third priority is transfer. A method should be re-evaluated across sites, devices, languages, organizations, or physical regimes before broad claims are made. For the present focus, domain-shift abstention therefore becomes a criterion for deciding which evidence deserves further scrutiny.

The field also needs infrastructure for cumulative, revisable evidence. Benchmarks should maintain negative results, failed branches, and changed definitions so later work does not learn only from successful demonstrations. Shared reporting

should include data lineage, versioned testing code, confidence intervals, and enough error material to reproduce major conclusions. Prospective studies should predefine monitoring and stopping rules, then measure how people adapt to the application. Finally, researchers should compare simple baselines with complex architectures under equivalent information. Complexity is justified when it adds a measurable capability - for example, structural localization, calibrated deferral, or incremental updating - not merely when it creates a richer narrative around the same errors. Seen through domain-shift abstention, the issue is not merely technical; it determines which claims remain open to challenge.

## Conclusion

Domain-shift abstention is credible only when it is attached to an documented evidence chain and decision policy. The review identifies several concrete mechanisms: structured exploration, typed graphs, conditional revision, adaptive computation, physical constraints, and repeated validation. At the same time, success on the originating benchmark cannot define a safe implementation boundary. For filings, news, and social media, the practical standard should be a traceable pipeline that measures shift and instability, supports selective revision, distillation, abstention, expert review, and continuous validation, and preserves enough evidence for an independent reviewer to reconstruct failure. Future work should compare safeguards under realistic resource and staffing limits, publish negative and subgroup results, and treat monitors and automated judges as components requiring their own validation. A dependable system need not answer every case; it must connect each action to supporting evidence, respond appropriately to uncertainty, and leave an auditable record. Seen through domain-shift abstention, the issue is not merely technical; it determines which claims remain open to challenge.

## References

1. Zhang, Yin, et al. "SAINF: Intrinsic Self-Correction for Robust Machine Translation with Large Language Models." Frontiers of Computer Science (2026).

2. Li, Yuanhao, et al. "DRAFT-RL: Multi-Agent Chain-of-Draft Reasoning for Reinforcement Learning-Enhanced LLMs." Proceedings of the AAAI Conference on Artificial Intelligence 40.35 (2026): 29530-29537.

3. Zheng, Zihao, et al. "Reliable Financial Named Entity Recognition under Domain Shift." arXiv preprint arXiv:2608.19558 (2026).

4. Araci, Dogu. "FinBERT: Financial Sentiment Analysis with Pre-Trained Language Models." arXiv preprint arXiv:1908.10063, 2019.

5. Loughran, Tim, and Bill McDonald. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." Journal of Finance, vol. 66, no. 1, 2011, pp. 35-65.

6. Gururangan, Suchin, et al. "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8342-8360.

7. Vaswani, Ashish, et al. "Attention Is All You Need." Advances in Neural Information Processing Systems, vol. 30, 2017.

8. Sennrich, Rico, Barry Haddow, and Alexandra Birch. "Neural Machine Translation of Rare Words with Subword Units." Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016, pp. 1715-1725.

9. Papineni, Kishore, et al. "BLEU: A Method for Automatic Evaluation of Machine Translation." Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311-318.

10. Rei, Ricardo, et al. "COMET: A Neural Framework for MT Evaluation." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 2685-2702.

11. Wang, Xuezhi, et al. "Self-Consistency Improves Chain of Thought Reasoning in Language Models." International Conference on Learning Representations, 2023.

12. Ouyang, Long, et al. "Training Language Models to Follow Instructions with Human Feedback." Advances in Neural Information Processing Systems, vol. 35, 2022, pp. 27730-27744.

13. Rafailov, Rafael, et al. "Direct Preference Optimization: Your Language Model Is Secretly a Reward Model." Advances in Neural Information Processing Systems, vol. 36, 2023.