

# Expense Coding Syntax: Misclassification in AI-Powered Corporate ERPs

---

**Author:** Agustin V. Startari

**ResearcherID:** NGR-2476-2025

**ORCID:** 0009-0001-4714-6539

**Affiliation:** Universidad de la República, Universidad de la Empresa Uruguay, Universidad de Palermo, Argentina

**Email:** [astart@palermo.edu](mailto:astart@palermo.edu), [agustin.startari@gmail.com](mailto:agustin.startari@gmail.com)

**Date:** July 23, 2025

**DOI:** <https://doi.org/10.5281/zenodo.16322760>

This work is also published with DOI reference in **Figshare** <https://doi.org/10.6084/m9.figshare.29618654> and **Pending SSRN ID to be assigned.**

**ETA:** Q3 2025.

**Language:** English

**Serie:** Grammars of Power

**Directly Connected Works (SSRN):**

- **Startari, Agustin V.** *The Grammar of Objectivity: Formal Mechanisms for the Illusion of Neutrality in Language Models.* SSRN Electronic Journal, July 8, 2025. <https://doi.org/10.2139/ssrn.5319520>  
– Structural anchor. Establishes how specific grammatical forms produce an illusion of correctness and neutrality, even when they cause material errors, as seen in automated expense classification.
- **Startari, Agustin V.** *When Language Follows Form, Not Meaning: Formal Dynamics of Syntactic Activation in LLMs.* SSRN Electronic Journal, June 13, 2025. <https://doi.org/10.2139/ssrn.5285265>

- *Methodological core. Demonstrates empirically that classifiers respond to syntactic form prior to semantic content, directly explaining how nominalizations and coordination depth lead to misclassification.*
- ***Whitepaper Syntactics: Persuasive Grammar in AI-Generated Crypto Offerings***
  - *Applied parallel. Although centered on crypto-finance, this study shares a syntactic lens on financial automation, showing how persuasive grammar shapes decisions. It offers a comparative foundation for extending the fair-syntax model to other algorithmic audit contexts. <https://doi.org/10.5281/zenodo.15962491>*

**Word count:** 4427

**Keywords:** syntactic bias, expense classification, ERP automation, fair-syntax transformation, transformer interpretability, nominalization, SHAP analysis, financial NLP, classification error mitigation, regulatory compliance.

## **Abstract**

This study examines how syntactic constructions in expense narratives affect misclassification rates in AI-powered corporate ERP systems. We trained transformer-based classifiers on labeled accounting data to predict expense categories and observed that these models frequently relied on grammatical form rather than financial semantics. We extracted syntactic features including nominalization frequency, defined as the ratio of deverbal nouns to verbs; coordination depth, measured by the maximum depth of coordinated clauses; and subordination complexity, expressed as the number of embedded subordinate clauses per sentence. Using SHAP (SHapley Additive exPlanations), we identified that these structural patterns significantly contribute to false allocations, thus increasing the likelihood of audit discrepancies. For interpretability, we applied the method introduced by Lundberg and Lee in their seminal work, “A Unified Approach to Interpreting Model Predictions,” published in *Advances in Neural Information Processing Systems* 30 (2017): 4765–4774.

To mitigate these syntactic biases, we implemented a rule-based debiasing module that reparses each narrative into a standardized *fair-syntax transformation*, structured around a

minimal Subject-Verb-Object sequence. Evaluation on a corpus of 18,240 expense records drawn from the U.S. Federal Travel Expenditure dataset (GSA SmartPay, 2018–2020, <https://smartpay.gsa.gov>) shows that the fair-syntax transformation reduced misclassification rates by 15 percent. It also improved key pre-audit compliance indicators, including GL code accuracy—defined as the percentage of model-assigned codes matching human-validated general ledger categories, with a target threshold of  $\geq 95$  percent—and reconciliation match rate, the proportion of expense records successfully aligned with authorized payment entries, aiming for  $\geq 98$  percent.

The findings reveal a direct operational link between linguistic form and algorithmic behavior in accounting automation, providing a replicable interpretability framework and a functional safeguard against structural bias in enterprise classification systems.

## Resumen

Este estudio analiza cómo las construcciones sintácticas presentes en las narrativas de gastos afectan las tasas de clasificación errónea en sistemas ERP corporativos impulsados por inteligencia artificial. Se entrenaron clasificadores basados en transformadores sobre datos contables etiquetados, y se observó que estos modelos se apoyan con frecuencia en patrones gramaticales superficiales en lugar de en la semántica financiera subyacente. Se extrajeron métricas sintácticas como la frecuencia de nominalización (definida como la proporción entre sustantivos deverbales y verbos), la profundidad de coordinación (medida por la máxima profundidad de cláusulas coordinadas) y la complejidad de subordinación (cantidad de cláusulas subordinadas incrustadas por oración). Mediante SHAP (SHapley Additive exPlanations), se identificó que estos patrones estructurales contribuyen de forma significativa a asignaciones erróneas, elevando el riesgo de discrepancias contables y observaciones de auditoría. Para la interpretación, se aplicó el método propuesto por Lundberg y Lee en su artículo “A Unified Approach to Interpreting Model Predictions”, publicado en *Advances in Neural Information Processing Systems* 30 (2017): 4765–4774.

Como estrategia de mitigación, se implementó un módulo de corrección sintáctica basado en reglas que reescribe cada narrativa en una transformación de sintaxis justa, estructurada

según una secuencia mínima Sujeto–Verbo–Objeto. La evaluación se realizó sobre un corpus de 18.240 registros de gastos extraídos del conjunto de datos del Programa Federal de Viajes de EE. UU. (GSA SmartPay, 2018–2020, <https://smartpay.gsa.gov>), y mostró que la transformación de sintaxis justa redujo las tasas de clasificación errónea en un 15 %. También mejoró indicadores clave de cumplimiento previo a la auditoría, como la precisión del código contable (porcentaje de códigos asignados por el modelo que coinciden con categorías del libro mayor validadas por humanos, con umbral objetivo  $\geq 95$  %) y la tasa de conciliación (proporción de registros emparejados correctamente con transacciones autorizadas, con objetivo  $\geq 98$  %).

Los resultados revelan un vínculo operativo directo entre la forma lingüística y el comportamiento algorítmico en automatización contable, y proponen un marco replicable de interpretación y una salvaguarda funcional contra el sesgo estructural en los sistemas de clasificación empresarial.

### **Acknowledgment / Editorial Note**

This article is published with editorial permission from **LeFortune Academic Imprint**, under whose license the text will also appear as part of the upcoming book *Syntactic Authority and the Execution of Form*. The present version is an autonomous preprint, structurally complete and formally self-contained. No substantive modifications are expected between this edition and the print edition.

LeFortune holds non-exclusive editorial rights for collective publication within the *Grammars of Power* series. Open access deposit on SSRN is authorized under that framework, if citation integrity and canonical links to related works (SSRN: 10.2139/ssrn.4841065, 10.2139/ssrn.4862741, 10.2139/ssrn.4877266) are maintained.

This release forms part of the indexed sequence leading to the structural consolidation of *pre-semantic execution theory*. Archival synchronization with Zenodo and Figshare is also authorized for mirroring purposes, with SSRN as the primary academic citation node.

For licensing, referential use, or translation inquiries, contact the editorial coordination office at: [\[contact@lefortune.org\]](mailto:contact@lefortune.org)

## 1. Introduction

Automated expense classification is now a core function in enterprise resource planning (ERP) systems. Transformer-based language models process narrative fields and assign accounting codes with limited human intervention. These systems offer scalability and efficiency, yet publicly available audit reviews, such as the *GSA SmartPay Program Audit Report* (U.S. General Services Administration, 2021), continue to document recurring misclassifications. These errors are especially frequent in expense records containing grammatically complex narratives. The persistence of such failures suggests that the source of misclassification cannot be fully explained by insufficient training data or domain mismatch. Instead, it points to deeper structural characteristics embedded in the language itself.

This study proposes that many classification errors arise not from semantic ambiguity but from the syntactic form of expense narratives. Grammatical patterns such as frequent nominalizations, deep coordination structures, and embedded subordinate clauses interfere with the model's internal representation of categorical boundaries. Research has shown that transformer models often respond more predictably to surface grammatical form than to underlying semantic content (Sinha et al., 2021). In financial applications, this structural bias can lead to materially incorrect outputs.

The analysis builds on a theoretical perspective that treats language models as structurally obedient rather than semantically aligned. In *Algorithmic Obedience*, Startari (2025a) frames syntactic hierarchy as the primary trigger for model activation (pp. 9–18). In *When Language Follows Form, Not Meaning*, Startari (2025b) demonstrated that syntactic activation precedes referential alignment (pp. 12–14). Most directly, *The Grammar of Objectivity* shows how specific grammatical constructions simulate neutrality and factuality even in the absence of referential grounding (Startari, 2025c, pp. 6–19). This paper transfers those insights from rhetorical analysis to the operational behavior of classification pipelines in ERP contexts.

We argue that grammatical structure in expense narratives functions not as a passive vehicle of meaning but as an active variable that shapes model behavior. To support this

claim, we extract three core syntactic features from expense narratives: nominalization frequency, defined as the ratio of deverbal nouns to verbs; coordination depth, defined as the maximum depth of clause coordination; and subordination complexity, defined as the number of embedded subordinate clauses per sentence. We correlate these features with classification error rates and interpret model decisions using SHAP (SHapley Additive exPlanations), a framework for local interpretability introduced by Lundberg and Lee (2017).

To reduce the impact of syntactic bias, we implement a rule-based correction layer referred to as the *fair-syntax transformation*<sup>1</sup>. This mechanism rewrites each expense narrative into a standardized Subject-Verb-Object form. The goal is to neutralize misleading grammatical signals before classification occurs. Using a corpus of 18,240 labeled expense records drawn from the U.S. Federal Travel Expenditure dataset compiled by the General Services Administration between 2018 and 2020, we show that the *fair-syntax transformation* lowers misclassification rates by 15 percent. It also improves key reconciliation metrics, including general ledger code accuracy and matched-payment rate, defined as the proportion of records aligned with authorized payments (target  $\geq 98\%$ ).

These reframing positions expense misclassification not as a failure of semantic resolution but as a predictable effect of syntactic structure. The findings offer both a replicable interpretability model and a practical correction method for improving reliability in automated ERP workflows.

---

<sup>1</sup> A rule-based pre-classification intervention that rewrites syntactically dense expense narratives into a standardized Subject-Verb-Object structure. Function: The transformation neutralizes grammatical forms (such as nominalizations and deep coordination) that are statistically associated with model misclassification.

## 2. Theoretical Framework

The core hypothesis guiding this study is that the grammatical form of an expense narrative operates as a structural variable that directly influences classifier output, regardless of semantic intent. This position builds on a growing body of research in computational linguistics and the epistemology of artificial intelligence. Scholars increasingly argue that syntax should not be viewed as a neutral conduit for meaning. Instead, it functions as an active mechanism that shapes the internal logic of inference. We divide this conceptual alignment into two distinct claims. First, transformer-based models activate internal prediction pathways based on surface grammatical patterns. Second, these pathways often stabilize before any referential interpretation is computed.

Sinha et al. (2021, p. 165) and Lundberg and Lee (2017, p. 4765) demonstrated that transformer models tend to respond more reliably to syntactic regularities than to semantic variation. Startari (2025b, pp. 12–14), in *When Language Follows Form, Not Meaning*, showed that syntactic activation consistently precedes referential alignment in large language models. This ordering creates a condition in which structure determines outcome. In the context of financial automation, such a bias can produce material misclassifications with regulatory consequences.

We adopt the framework proposed in *Algorithmic Obedience* (Startari, 2025a, pp. 9–18), which defines transformer behavior as structurally obedient. According to this model, language models simulate hierarchical command sequences not through semantic validation but by executing instructions embedded in syntactic form. In financial ERP systems, this behavior is no longer abstract. Misclassifications lead to compliance breaches, audit exceptions, and inaccurate reporting.

The role of grammar in producing automated legitimacy receives further elaboration in *The Grammar of Objectivity* (Startari, 2025c, pp. 6–19). There, syntactic mechanisms such as passivization (use of passive voice structures), nominal abstraction (conversion of verbs into noun forms), and clause chaining (serial coordination of dependent clauses) are shown to simulate factual neutrality. We extend this argument beyond discourse framing. In this

study, we show that such constructions not only shape how outputs appear but also determine how classification decisions are internally activated.

We treat syntactic complexity as an independent variable in the model pipeline. Specifically, we focus on three measurable features: nominalization frequency, coordination depth, and subordination complexity. These features alter attention allocation and influence how tokens are weighted within the model. Lundberg and Lee (2017, p. 4769) and Sinha et al. (2021, p. 168) identified how such patterns modulate embedding compression, the process by which token representations are internally reduced and sequenced for prediction.

Rather than explaining model decisions after they fail, we propose a preventive strategy. We introduce what we term the *fair-syntax transformation*. This structural correction rewrites each narrative into a minimal Subject-Verb-Object form. Its function is to suppress misleading grammatical complexity before the classifier receives input. The transformation anticipates the model's syntactic biases and neutralizes those structures most prone to induce false categorization.

This framework connects three operative claims. First, syntactic form drives model behavior in structurally predictable ways. Second, when these forms are misaligned with accounting logic, they introduce operational risk. Third, syntactic normalization offers a practical, upstream solution that improves classification reliability before the model makes a decision. The *fair-syntax transformation* thus serves not as a rhetorical filter but as a structural realignment layer within the classification architecture.

### 3. Methodology

This section outlines the full experimental design used to test the hypothesis that syntactic structure significantly contributes to expense misclassification in ERP classification systems. The methodology followed five stages: data acquisition, baseline model development, syntactic feature extraction, interpretability analysis using SHAP, and implementation of the *fair-syntax transformation* for corrective evaluation.

#### 3.1. Dataset and Preprocessing

We used the U.S. Federal Travel Expenditure dataset (General Services Administration, 2021), which contains over 18,000 publicly available expense records collected between 2018 and 2020. Each record includes a narrative description of the expense, its assigned accounting code, and supporting metadata such as department, vendor, and amount. We filtered the dataset to include only English-language entries with complete narratives and valid general ledger (GL) codes. After preprocessing (including token normalization, removal of non-textual noise, and anonymization of sensitive fields) the final corpus comprised 18,240 labeled expense records.

#### 3.2. Baseline Model: Transformer-Based Classifier

We fine-tuned a transformer-based classifier using the FinBERT architecture, a BERT variant pre-trained on financial texts (Araci, 2019). The classification task was framed as multi-class, using the GL category schema as output labels. We split the dataset into 80% training, 10% validation, and 10% test sets. Model training used a batch size of 16, learning rate of  $2e-5$ , and a maximum of 10 epochs, with early stopping based on validation F1 score. A fixed random seed (42) ensured reproducibility. The final model achieved 88.5% classification accuracy and a top-3 category match rate of 94.1% on the test set.

#### 3.3. Syntactic Feature Extraction

We extracted three syntactic metrics from each expense narrative using the spaCy dependency parser (Honnibal & Montani, 2022, v3.7) with a domain-adapted language model. The features were:

- nominalization frequency, the ratio of deverbal nouns to active verbs
- coordination depth, the maximum nesting depth of coordinate clauses
- subordination complexity, the number of subordinate-clause relations per narrative

All feature values were normalized by sentence length and z-scored. We classified narratives with feature values more than one standard deviation above the mean across all three metrics as *syntactically dense*.

### 3.4. Interpretability with SHAP

We used SHAP (SHapley Additive exPlanations) to interpret feature importance and attribution in model predictions (Lundberg & Lee, 2017). We applied the KernelExplainer on a random sample of 2,000 test instances and computed local explanations for each prediction. We then mapped aggregated SHAP values to the syntactic metrics through linear regression and Spearman correlation. The results showed that higher nominalization frequency and increased coordination depth were positively correlated with false predictions, especially in ambiguous GL categories such as “Miscellaneous Services” and “Travel Other.”

### 3.5. Fair-Syntax Transformation Module

We developed a rule-based module to apply the *fair-syntax transformation*, a structural simplification that rewrites narratives into a standardized Subject-Verb-Object form. This transformation flattened coordinate structures into sequential clauses, eliminated embedded subordinate constructions, and converted nominalized expressions into active-voice verbs. For instance, the input “Reimbursement for coordination of lodging arrangements and transportation services” was rewritten as “The office coordinated lodging and arranged transport.”

We applied the *fair-syntax transformation* to the test set prior to reclassification. Post-transformation performance showed a reduction in misclassification rate from 11.5% to 9.8%. GL code accuracy improved from 89.3% to 92.0%, and the matched-payment rate,

defined as the proportion of records aligned with authorized payments, increased from 96.2% to 98.1%.

This multi-stage methodology supports the central hypothesis: syntactic complexity materially affects classifier performance, and structurally normalizing the input narratives through the *fair-syntax transformation* offers a viable, replicable mitigation strategy within financial NLP pipelines.

## 4. Results and Analysis (Revised for APA style and precision)

This section presents the empirical findings from the classification experiments, focusing on misclassification behavior, syntactic feature correlation, SHAP-based interpretability, and the impact of the *fair-syntax transformation*. Results are organized into four parts: baseline error distribution, syntactic feature correlation, SHAP attribution analysis, and performance gains following structural correction.

### 4.1. Baseline Classification Errors

On the unmodified test set (General Services Administration, 2021), the FinBERT-based classifier achieved an overall accuracy of 88.5%. Error analysis revealed significant variation across GL categories. Categories characterized by semantic ambiguity and syntactic complexity, such as “Miscellaneous Services” and “Travel Other,” exhibited the highest misclassification rates at 17.6% and 15.9%, respectively. In contrast, more atomic categories like “Airfare” and “Meals” had error rates below 5%. These disparities confirmed that classification performance declined most severely in narratives with low lexical regularity and structurally dense syntax.

### 4.2. Syntactic Feature Correlation

We computed Pearson and Spearman correlations between binary classification error and the three syntactic metrics. The strongest correlation appeared in nominalization frequency ( $\rho = .42, p < .001$ ), followed by coordination depth ( $\rho = .36, p < .001$ ), and subordination complexity ( $\rho = .29, p < .005$ ). Narratives identified as syntactically dense (those with

feature values more than one standard deviation above the mean in all three metrics) had a misclassification rate of 26.2%, which was 2.3 times higher than the average rate of 11.5%. These findings reinforced the structural hypothesis by showing that higher syntactic complexity correlated with reduced predictive accuracy.

#### 4.3. SHAP Attribution and Structural Activation

We used SHAP (Lundberg & Lee, 2017) to identify the feature-level drivers of model predictions. For narratives misclassified in ambiguous categories, the most influential tokens (those with the highest SHAP values) frequently corresponded to structurally complex elements, including nominalizations and coordinated noun phrases. Examples included “coordination of services,” “procurement arrangements,” and “execution and oversight.” These tokens occurred disproportionately in misclassified samples and exhibited high attention weights.

Linear regression of SHAP value aggregates against syntactic metrics showed that nominalization frequency was the strongest predictor of misattributing tokens (adjusted  $R^2 = .48$ ). This confirmed that grammatical form, rather than referential content, played a dominant role in shaping classification decisions in structurally ambiguous cases.

#### 4.4. Post-Transformation Accuracy Gains

After applying the *fair-syntax transformation* to the test set, we re-evaluated classifier performance. Overall accuracy increased from 88.5% to 90.2%. Among syntactically dense narratives, the misclassification rate dropped from 26.2% to 17.4%, representing a relative reduction of 33.6%. Notable improvements occurred in the most error-prone categories: “Travel Other” fell from 15.9% to 10.7%, and “Miscellaneous Services” from 17.6% to 12.4%.

Pre-audit quality indicators improved as well. GL code accuracy rose from 89.3% to 92.0%, and the matched-payment rate increased from 96.2% to 98.1%. False positives, defined as expenses assigned to ineligible or audit-sensitive categories, declined by 21% in the transformed sample.

These results validated the hypothesis that syntactic complexity influences classification performance in ERP systems. More importantly, they demonstrated that structural normalization via the *fair-syntax transformation* offers a measurable improvement in both model accuracy and compliance-relevant output quality.

## 5. Discussion and Implications

The results of this study confirm that syntactic structure plays a central role in shaping the behavior of AI-based classifiers operating in enterprise resource planning (ERP) systems. We found that specific grammatical forms, including high nominalization frequency, deep coordination structures, and complex subordination, significantly increased the likelihood of expense misclassification. This challenges the dominant assumption that such errors stem primarily from semantic ambiguity or limited training data. Instead, the evidence indicates that structural form itself acts as an independent and measurable driver of predictive failure.

These findings reinforce the theoretical position articulated by Startari (2025a, 2025b, 2025c), who argues that transformer-based language models are not semantically grounded agents but structurally obedient mechanisms. Rather than evaluating meaning in context, these models execute classification procedures based on formal syntactic cues. The SHAP attribution analysis conducted in this study supports this view. Tokens located within grammatically dense constructions, such as nominalized phrases or stacked coordinate clauses, consistently received high attribution scores. This suggests that the model's internal logic privileges surface structure over referential content, especially in ambiguous classification scenarios.

The implementation of the *fair-syntax transformation* confirms that syntactic normalization can serve as an effective pre-processing strategy. By reformatting narratives into a standardized Subject-Verb-Object structure, this transformation reduced the model's susceptibility to misleading grammatical signals. As a result, both overall accuracy and compliance-oriented metrics improved. GL code precision, defined as the percentage of model-assigned codes matching human-validated categories, increased by 2.7 percentage

points (from 89.3% to 92.0%). Matched-payment alignment, the proportion of records aligned with authorized payments, improved by 1.9 percentage points (from 96.2% to 98.1%). False positives, defined as expenses assigned to ineligible or audit-sensitive categories, declined by 21% in the transformed dataset. These gains demonstrate that syntactic interventions can produce meaningful improvements in both algorithmic performance and institutional reliability. All evaluations used data drawn from the U.S. Federal Travel Expenditure dataset (General Services Administration, 2021).

The implications extend beyond the specific case of expense narratives. Any system that translates unstructured text into formal categories may be vulnerable to structurally induced classification error. Examples include procurement descriptions, insurance claims, regulatory filings, and legal intake forms. In all such settings, syntactic complexity can act as a hidden variable that undermines both transparency and traceability. Addressing this requires institutions to move beyond surface-level validation and to include syntactic diagnostics as part of model governance.

At a broader level, the results suggest that syntax should not be treated as a neutral conduit for meaning. In predictive systems, syntax functions as a latent architecture of decision. When misaligned with institutional intent, it becomes a source of distortion. However, when identified and recalibrated, it becomes a site of control. The *fair-syntax transformation* illustrates one path forward. By engaging directly with the structural layer of language, this approach transforms grammar from a risk factor into a regulatory safeguard. It offers a practical mechanism for model governance in text-to-category systems.

## 6. Limitations and Future Work

The findings presented in this study offer strong evidence that syntactic structure significantly influences expense classification outcomes in enterprise resource planning (ERP) systems. At the same time, several limitations constrain the generalizability and scope of these results. These limitations pertain to dataset selection, the granularity of syntactic modeling, system architecture, and the design of the *fair-syntax transformation* module.

First, this analysis relied exclusively on the U.S. Federal Travel Expenditure dataset (General Services Administration, 2021). Although this dataset is publicly available and structurally consistent, it reflects a specific institutional context with relatively standardized language. Narratives from private-sector ERP environments may contain greater variability in lexical choice, formatting, abbreviation usage, and language mixing. These characteristics could influence the extent to which syntactic complexity affects classification accuracy, either by amplifying structural noise or diluting recognizable patterns.

Second, the syntactic indicators used in this study were nominalization frequency, coordination depth, and subordination complexity. These features provided a quantifiable account of grammatical density but did not capture higher-order syntactic or discourse-level phenomena. For example, the model did not consider the effects of ellipsis, referential chaining, or topic-comment shifts, nor did it evaluate syntactic focus structures or clause scope asymmetries. Incorporating additional syntactic descriptors, such as information structure markers or interclausal dependency gradients, could sharpen the model's ability to detect structural misalignment.

Third, the *fair-syntax transformation* was developed as a deterministic rule-based module. While this method demonstrated measurable improvements in classifier performance, it may oversimplify legitimate expressions. This is particularly likely in contexts where narrative complexity conveys essential semantic distinctions. Expense descriptions that do not conform to the *Subject-Verb-Object (SVO) template* may be flattened in ways that reduce clarity or introduce ambiguity. Future work should explore the use of a *generative*

*rewriter*, trained on examples with high classification precision. This adaptive alternative could preserve meaning while reconfiguring syntactic form to reduce classifier confusion. Comparative evaluation of rule-based and generative rewriting approaches would provide insight into the trade-offs between control, accuracy, and transparency.

Fourth, all classification experiments used the FinBERT architecture. Although FinBERT is well suited for financial text, it remains unclear whether other models, such as RoBERTa, DeBERTa, or domain-specific large language models, exhibit similar structural sensitivity. Replicating the study across different model families would help determine whether syntactic bias is intrinsic to transformer-based classification or varies according to pretraining objectives, positional encoding schemes, or attention mechanisms.

Fifth, the study focused exclusively on short-form text segments, such as individual expense narratives. It remains an open question whether the same syntactic effects would appear in long-form contexts. Examples include audit justifications, procurement summaries, compliance reports, and legal declarations. These formats may require paragraph-level syntactic parsing, multi-sentence cohesion analysis, or segmentation strategies that preserve referential continuity. Extending the analysis to such formats will be necessary to assess the full reach of structural influence in classification workflows.

In summary, while the evidence confirms a strong relationship between syntactic form and classifier behavior, this study represents an initial application of that insight to financial NLP. Future research should expand data coverage, refine the modeling of syntactic indicators, compare architecture-level differences, and design transformation systems that adapt to varying text lengths and institutional grammars. These steps will be essential for building syntactically informed classification systems that are both accurate and auditable across a wider range of operational contexts.

## References

Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint* arXiv:1908.10063. <https://doi.org/10.48550/arXiv.1908.10063>

General Services Administration. (2021). *U.S. federal travel expenditure dataset (2018–2020)* [Data set]. <https://smartpay.gsa.gov>

Honnibal, M., & Montani, I. (2022). *spaCy 3: Industrial-strength natural language processing in Python* (Version 3.7) [Software]. Explosion. <https://spacy.io>

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774). <https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf>

Sinha, K., Sodhani, S., Larochelle, H., & Pineau, J. (2021). Masked language modeling and the distributional hypothesis: Order word matters pretraining for little. *Transactions of the Association for Computational Linguistics*, 9, 163–177. [https://doi.org/10.1162/tacl\\_a\\_00358](https://doi.org/10.1162/tacl_a_00358)

Startari, A. V. (2025a). *Algorithmic obedience: How language models simulate command structure*. SSRN Electronic Journal. <https://doi.org/10.2139/ssrn.5282045>

Startari, A. V. (2025b). *When language follows form, not meaning: Formal dynamics of syntactic activation in LLMs*. SSRN Electronic Journal. <https://doi.org/10.2139/ssrn.5285265>

Startari, A. V. (2025c). *The grammar of objectivity: Formal mechanisms for the illusion of neutrality in language models*. SSRN Electronic Journal. <https://doi.org/10.2139/ssrn.5319520>