Modeling Professionalism in Expert Questioning through Linguistic Differentiation

Giulia D'Agostino
Università della Svizzera italiana,
Switzerland
dagosgi@usi.ch

Chung-Chi Chen
National Institute of Advanced
Industrial Science and Technology,
Japan
c.c.chen@acm.org

Abstract

Professionalism is a crucial yet underexplored dimension of expert communication, particularly in high-stakes domains like finance. This paper investigates how linguistic features can be leveraged to model and evaluate professionalism in expert questioning. We introduce a novel annotation framework to quantify structural and pragmatic elements in financial analyst questions, such as discourse regulators, prefaces, and request types. Using both human-authored and large language model (LLM)-generated questions, we construct two datasets: one annotated for perceived professionalism and one labeled by question origin. We show that the same linguistic features correlate strongly with both human judgments and authorship origin, suggesting a shared stylistic foundation. Furthermore, a classifier trained solely on these interpretable features outperforms gemini-2.0 and SVM baselines in distinguishing expert-authored questions. Our findings demonstrate that professionalism is a learnable, domain-general construct that can be captured through linguistically grounded modeling.

1 Introduction

Professionalism is a vital yet underdefined dimension of language use in high-stakes communicative settings. Whether in corporate finance, medical consultations, legal interviews, or hiring processes, expert questioning demands more than fluency or grammatical correctness—it requires clarity, relevance, and strategic pragmatics. Despite its intuitive importance, modeling professionalism in language remains elusive, in part due to the lack of consistent, operational definitions and annotated datasets. Current NLP systems excel at generating well-formed text but often struggle with the subtleties of professional discourse. In expert interactions, subtle linguistic cues, such as how a question is framed, how background is contextualized, or how politeness is conveyed, can determine whether an utterance is perceived as informed and effective or vague and unprofessional. These micro-level elements are difficult to define explicitly, but they are deeply rooted in linguistic structure.

We argue that such linguistically grounded features offer a promising path forward. Rather than relying solely on surface metrics like sentence length or lexical complexity, we focus on structural and pragmatic properties of questions: the use of prefaces, the types of requests embedded, the presence of discourse regulators, and more. These elements reflect not just what is being asked, but how the speaker positions themselves in relation to the audience, the topic, and the communicative norms of the domain. In this work, we explore whether such features can be operationalized to distinguish between more and less professional forms of expert questioning. We develop an annotation pipeline to extract and quantify a range of linguistically motivated features, and we investigate how these features correlate with human perceptions of professionalism.

To evaluate their predictive power, we introduce a controlled classification task. Specifically, we examine whether questions authored by human experts can be reliably distinguished from those generated by large language models (LLMs), under matched conditions. While our ultimate goal is to model professionalism, this origin-based contrast serves as a proxy—allowing us to isolate the stylistic and pragmatic signatures of expert discourse in a structured, observable way. Our results show that linguistic features not only support accurate origin classification, but also align strongly with human judgments of professionalism. This provides both a practical tool for evaluating generated language and a theoretical foundation for future work on modeling professionalism in a domain-general, interpretable manner.

2 Related Work

Professional communication differs from everyday language through its use of formal, goal-oriented, and domain-specific features. Prior work shows that institutional discourse is shaped by roles and communicative goals, producing distinctive linguistic patterns (Drew and Heritage, 1992). Across domains like medicine and law, professionals use technical terms and polite, precise language to convey expertise (Sarangi and Roberts, 1999). These features serve functional roles—ensuring clarity, maintaining face, and aligning with institutional norms. Our work builds on this foundation by treating professionalism as a measurable linguistic trait in expert communication.

Questions in expert settings often serve strategic purposes beyond information seeking. Studies highlight how their design—yes/no vs. wh-, direct vs. indirect—can signal stance and control responses (Clayman and Heritage, 2002). For example, adversarial questions can exert pressure or challenge authority (Clayman et al., 2007). Professionals across domains similarly embed rhetorical aims in question form, balancing politeness with critical inquiry (Ilie, 2021). Our approach draws on these findings to identify linguistic markers of professional questioning.

LLM outputs differ systematically from human writing, even when fluent. Corpus studies show LLMs tend to produce more uniform, less varied, and less contextually rich language (Berber Sardinha, 2024; Zanotto and Aroyehun, 2024). These differences inform detection methods and motivate our work: we focus on the nuanced features of professional human questioning that current models struggle to replicate. Rather than pure classification, we analyze interpretable linguistic indicators of professionalism in human vs. AI-generated questions.

Earnings calls are a prime domain for studying expert dialogue. These events feature formal yet interactive Q&A between executives and analysts, marked by strategic questioning within professional constraints (Crawford Camiciottoli, 2010; de Oliveira and Pereira, 2018). Analysts often use indirect or framed questions to press for clarity while maintaining collegiality. Language choices in this context can impact perceptions and market outcomes. Our work extends prior studies by computationally modeling professionalism in analyst questions, using human-AI differences as a lens to capture these expert traits (Palmieri et al., 2015; D'Agostino et al., 2024).

3 Connecting Professionalism and Question Origin

In this section, we explore the relationship between perceived professionalism and question origin. While our ultimate goal is to model professionalism as it is perceived by human evaluators, directly learning such a construct is challenging due to the subjectivity and variability of human judgments. Instead, we propose that comparing questions authored by domain experts to those generated by large language models (LLMs) offers a viable contrastive setting that surfaces the linguistic cues associated with professional communication.

We present three empirical components supporting this claim. First, we introduce the datasets used in our study, including a crowdsourced annotation of professionalism scores. Second, we describe the linguistically motivated features used to characterize each question. Finally, we analyze how these features correlate with both human professionalism ratings and question origin, showing that the same stylistic and pragmatic traits are salient in both tasks.

3.1 Datasets and Annotation

To investigate professionalism in expert questioning, we construct two primary datasets. The first, which we refer to as the Human-Rated Professionalism Dataset (HRPD), consists of 250 questions sampled from earnings call Q&A sessions, equally balanced between those asked by financial analysts and those generated by LLMs. Each question was rated by five independent annotators on a 3-point scale for perceived professionalism in previous work by Juan et al. (2025).[1] This dataset provides a distribution of subjective human judgments, enabling us to examine which linguistic features align with perceived professionalism.

The second dataset, which we term the Question Origin Dataset (QOD), combines a separate set of 250 analyst questions and 250 LLM-generated questions, sampled from a broader pool but without human rating. This dataset is used for training and evaluating classification models on the binary task of determining whether a question is of human or machine origin. While the labels in this dataset reflect authorship, not professionalism per se, we

[1] Juan et al. (2025) shared the dataset upon request.

treat this task as a structured proxy to study how professional linguistic traits manifest at scale.

3.2 Linguistic Features

In order to characterize the pragmatic and structural properties of questions within financial dialogue, we rely on a linguistically-motivated feature set. These features are selected for their theoretical grounding in discourse analysis, pragmatics, and question typology, and they allow us to move beyond surface-level NLP metrics to a more interpretable, fine-grained representation of question design. We divide them into four principal categories: discourse regulators, prefaces, question types, and request types.

Discourse regulators are dialogical or meta-discursive components that organize the interactional flow of the question turn. They may acknowledge the previous speaker, specify who the question is for, indicate the number or order of questions, or introduce the topic explicitly. For example, expressions such as “Thanks,” “Chris, this one is for you,” or “A quick one on margins” serve to mediate the exchange in ways that are socially and structurally relevant. These elements are key in professional financial exchanges, where clarity, politeness, and sequential control are valued.

Prefaces, by contrast, are propositions that precede the question proper and function to justify, contextualize, or rationalize the upcoming interrogative act. They may state facts (“You said margins were up last quarter”), express opinions (“It seems like guidance is too optimistic”), report others’ speech (“As the CEO mentioned in the last call...”), or include meta-commentary (“I ask this because it wasn’t clear from the disclosure”). These reveal the epistemic stance of the speaker and are frequently used in professional contexts to construct credibility or soften face-threatening acts.

We also categorize the syntactic form of the question into three types: open (wh-questions), polar (yes/no), and closed-list (multiple-choice). These forms determine the amount and specificity of information requested and reflect the speaker’s strategic intent. Finally, we annotate the request type encoded in the question, that is, the kind of answer being sought—ranging from clarification and confirmation to explanations, data, or opinions. This level of pragmatic annotation allows us to approximate the functional load of each question.

To determine how these linguistic features align with perceived professionalism, we compute Spearman correlations between each feature and the professionalism scores gathered via questionnaire. Table 1 presents all statistically significant correlations in both the *mixed* and *combined* datasets, with positively correlated features listed first (top-down) and negatively correlated ones following.

	HRPD	QOD
Request types		
explanation	↑↑↑	↑
clarification	↓↓↓	↓↓
confirmation	↓↓↓	↓↓
Discourse regulators		
acknowledgment	↑↑↑	↓↓↓
recipient	↓↓↓	↓
theme	↓↓↓	↓↓↓
enumeration	↓↓↓	↓↓↓
counting	↓↓↓	↓↓↓
inside comment	↓↓↓	↓↓↓
Prefaces		
reported speech	↓↓	↓↓↓
opinion	↓↓↓	↓↓
fact	↓↓↓	↓↓↓
number	↓↓↓	↓↓↓
length	↓↓↓	↓↓↓
Question types		
open	↓↓↓	↑
polar	↓↓↓	↓↓↓
closed-list	↓↓↓	↓↓
NLP features		
type-token ratio	↑↑↑	↑↑↑
Flesch-Kincaid	↑↑↑	↑↑↑
Dale-Chall	↑↑↑	↑↑↑
word count	↓↓↓	↓↓↓
sentence count	↓↓↓	↓↓↓
NER (person) count	↓↓↓	↓↓
stopword count	↓↓↓	↓↓↓

Table 1: Spearman correlations with perceived professionalism. Arrows indicate direction and strength: ↑ = p < .05, ↑↑ = p < .01, ↑↑↑ = p < .001 (↓ for negative).

3.3 Converging Correlates

To assess whether perceived professionalism and question origin share common linguistic foundations, we compute Spearman correlations between the annotated features and two target variables:

1. **Perceived Professionalism:** The mean human rating (1–3) from the Human-Rated Professionalism Dataset.

2. **Question Origin:** A binary label (human or

LLM-generated) from the Question Origin Dataset.

Table 1 presents the features that are significantly correlated with both variables across the two datasets. We observe that many of the same linguistic indicators—such as preface frequency and length, request and question types, discourse structure, and readability—exhibit consistent correlation direction and strength in both tasks. For example, longer or more numerous prefaces correlate negatively with professionalism ratings and are more prevalent in machine-generated questions. Conversely, questions with clearer thematic focus and higher readability scores tend to be both rated as more professional and identified as human-authored. This convergence suggests that the linguistic traits perceived by humans as hallmarks of professionalism are the same traits that distinguish human expert questioning from LLM-generated imitations. We interpret this alignment as evidence that the question origin task can serve as a valid, linguistically grounded proxy for modeling professionalism, particularly in the absence of reliable large-scale annotation.

Interestingly, our results also challenge the common intuition that professional language must be complex or dense. In contrast, we find that higher professionalism ratings are associated with greater readability, fewer words and sentences, and more concise structure overall. This suggests that perceived professionalism may align more strongly with clarity and rhetorical control than with verbosity or technical elaboration. Finally, we observe higher variability in human ratings for naturally occurring questions, possibly due to their syntactic or pragmatic complexity. This further supports our use of question origin as a stable proxy for initial modeling, while acknowledging that true professionalism is inherently gradient and context-sensitive.

4 Predicting Professionalism with Linguistic Features

To evaluate whether linguistic features alone can predict professionalism, we train a Random Forest classifier using the full set of annotated features identified in earlier sections. We compare this model against two baselines:

- **SVM:** A support vector machine trained on raw question text using TF-IDF features.
- **gemini-2.0-flash (few-shot):** A prompt-based classifier using in-context examples, asked to predict whether a question was authored by a human or generated by a model.

Model	**Accuracy**	$\mathbf{F_1}$
Gemini	0.89	0.89
SVM	0.92	0.92
Random Forest	**0.96**	**0.96**

Table 2: Classification accuracy and F_1 score on the full question set. Our feature-based model outperforms both the SVM baseline and gemini-2.0-flash.

Table 2 reports accuracy and F_1 scores for the full classification task. The Random Forest model, despite using no text input, outperforms both baselines.

These results offer two key insights. First, the linguistic features identified in earlier analysis are not only correlated with professionalism, but are also predictive. Their discriminative power is sufficient to classify questions as human- or model-authored with high accuracy, surpassing even a large, general-purpose language model using in-context learning. Second, the fact that a shallow model using explicitly defined, interpretable features outperforms a black-box LLM suggests that professionalism is not an emergent property requiring deep modeling, but rather one that is encoded in recognizable and theoretically grounded linguistic signals. This supports our overall argument: that linguistic theory can guide effective, transparent modeling of communicative quality, especially in domains where clarity, etiquette, and rhetorical function matter.

5 Conclusion

We present a linguistically grounded approach to modeling professionalism in expert questioning. By analyzing structural and pragmatic features, we show that the same cues align with both human judgments and question origin. Our findings challenge the notion that complexity signals professionalism, and show that concise, readable questions are rated more professional. A lightweight classifier using these features outperforms both SVM and gemini-2.0-flash, demonstrating that professionalism is not only learnable, but also expressible through interpretable linguistic signals.

Limitations

Our study focuses on a single domain—financial earnings calls—which may limit generalizability to other professional contexts such as law or healthcare. While we treat question origin as a proxy for professionalism, this distinction cannot fully capture the gradient and context-sensitive nature of the construct. Moreover, although our crowdsourced annotations offer useful insights, variability in human ratings suggests that perceptions of professionalism are not uniform and may depend on factors beyond linguistic form.

Supplementary materials availability statement

The Question Origin Dataset and the test dataset for prediction evaluation are available at https://github.com/dagosgi/question-origin_professionalism_short/tree/main/supplementary-materials.

References

Tony Berber Sardinha. 2024. Ai-generated vs human-authored texts: A multidimensional comparison. *Applied Corpus Linguistics*, 4(1):100083.

Steven E. Clayman and John Heritage. 2002. *The News Interview: Journalists and Public Figures on the Air*. Cambridge University Press.

Steven E. Clayman, John Heritage, Marc N. Elliott, and Laurie L. McDonald. 2007. When does the watchdog bark? conditions of aggressive questioning in presidential news conferences. *American Sociological Review*, 72(1):23–41.

Belinda Crawford Camiciottoli. 2010. Earnings calls: Exploring an emerging financial reporting genre. *Discourse & Communication*, 4(4):343–359.

Giulia D'Agostino, Andrea Rocci, and Chris Reed. 2024. Capturing Analysts' Questioning Strategies in Earnings Calls via a Question Cornering Score (QCS). In *Proceedings of the Eighth Financial Technology and Natural Language Processing and the 1st Agent AI for Scenario Planning*, pages 107–118, Jeju, South Korea.

Maria L. de Oliveira and Sílvia M. R. Pereira. 2018. Formulations in delicate actions: A study of analyst questions in earnings conference calls. *International Journal of Business Communication*, 55(3):293–309.

Paul Drew and John Heritage. 1992. Analyzing talk at work: an introduction. In Paul Drew and John Heritage, editors, *Talk at Work: Interaction in Institutional Settings*, pages 3–65. Cambridge University Press.

Cornelia Ilie. 2021. Questions we (inter)act with: Interrelatedness of questions and answers in discourse. In Cornelia Ilie, editor, *Questioning and Answering Practices across Contexts and Cultures*, pages 1–32. John Benjamins.

Yining Juan, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2025. Co-trained retriever-generator framework for question generation in earnings calls. In *Companion Proceedings of the ACM Web Conference 2025*.

Rudi Palmieri, Andrea Rocci, and Nadzeya Kudrautsava. 2015. Argumentation in earnings conference calls. Corporate standpoints and analysts' challenges. *Studies in communication sciences*, 15, 2015(1):120–132.

Srikant Sarangi and Celia Roberts. 1999. *Talk, Work and Institutional Order: Discourse in Medical, Mediation and Management Settings*. Mouton de Gruyter, Berlin.

Silvia E. Zanotto and Segun Aroyehun. 2024. Human variability vs. machine consistency: A linguistic analysis of texts generated by humans and large language models. https://arxiv.org/abs/2412.03025. ArXiv:2412.03025.

A All Linguistic Features

Table 3 shows all features we used in the experiments.

B Use of AI Assistance in Writing

AI language tools were utilized to enhance grammar and phrasing during the writing process. These tools were solely employed for linguistic refinement; all aspects of research design, analysis, and content development were independently conducted and verified by the research team.

	HRPD	QOD
Request types		
explanation	↑↑↑	↑
clarification	↓↓↓	↓↓
confirmation	↓↓↓	↓↓
Discourse regulators		
acknowledgment	↑↑↑	↓↓↓
recipient	↓↓↓	↓
theme	↓↓↓	↓↓↓
enumeration	↓↓↓	↓↓↓
counting	↓↓↓	↓↓↓
inside comment	↓↓↓	↓↓↓
Prefaces		
reported speech	↓↓	↓↓↓
opinion	↓↓↓	↓↓
fact	↓↓↓	↓↓↓
number	↓↓↓	↓↓↓
length	↓↓↓	↓↓↓
Question types		
open	↓↓↓	↑
polar	↓↓↓	↓↓↓
closed-list	↓↓↓	↓↓
NLP features		
type-token ratio	↑↑↑	↑↑↑
Flesch-Kincaid	↑↑↑	↑↑↑
Dale-Chall	↑↑↑	↑↑↑
interjection count	↑↑↑	↓↓↓
word count	↓↓↓	↓↓↓
sentence count	↓↓↓	↓↓↓
filler word count	↓↓↓	↓↓↓
stopword count	↓↓↓	↓↓↓
NER (person) count	↓↓↓	↓↓
question count	↓↓↓	↓↓
assertion count	↓↓↓	↓↓↓
mean assertion length	↓↓↓	↓↓↓

Table 3: Spearman correlations with perceived professionalism. Arrows indicate direction and strength: ↑ = p < .05, ↑↑ = p < .01, ↑↑↑ = p < .001 (↓ for negative).