

Perceived Professionalism in Financial Inquiry: An Interpretable and Empirically Grounded Linguistic Modeling of Textual Preference

Giulia D’Agostino

Fraunhofer FKIE

Wachtberg, Germany

giulia.dagostino@fkie-extern.fraunhofer.de

Chung-Chi Chen

Human-Agent Ally Lab (HAA Lab)

National Institute of Informatics, Japan

chen@nii.ac.jp

Abstract

Professionalism is a central yet underdefined aspect of expert communication. This study investigates whether perceived professionalism can be modeled computationally in financial Q&A contexts, and whether non-expert preferences provide a meaningful basis for its annotation. We compile a mixed dataset of real and synthetically generated questions from earnings calls, annotated with human judgments of perceived professionalism and linguistically motivated features. Regression and classification experiments show that these features effectively capture evaluators’ preferences, enabling simple models to outperform both text-only and LLM baselines. The results demonstrate that interpretable, conceptually grounded linguistic features can rival more complex architectures while providing greater transparency, and support the value of linguistic awareness in NLP. More broadly, the study shows that modeling reader preferences can offer insights for the generation of professional questions.

1 Introduction

Professionalism is an evaluative parameter grounded in implicit prescriptive norms, whose descriptors remain largely undetermined. It can be either self-assessed or projected towards an external subject. It is often associated with socio-cultural parameters such as image (Kwon et al., 2025) or attitudes, behaviors, and values that are *expected* of a professional in a certain expert field (Bulk et al., 2019).

In this paper, we focus on *projected* professionalism applied to *communicative* practices; that is, the subjective evaluation made by the *recipient* of a communicative act produced in an expert domain. It reflects how well the communicative act aligns with the recipient’s expectations of clarity and effectiveness in a professional context. Effectively, we are investigating recipient preference alignment for inquiry formulation in a professional setting.

For short, this dimension will be simply called “professionalism” in the remainder of this contribution.

Despite its relevance to professional communication, professionalism remains underexplored as a linguistic construct, particularly with respect to its operationalization. Establishing explicit, interpretable descriptors for this notion could support applications within professional domains, such as the training of early-career managers, while also providing a basis for comparison across different expert contexts. At the same time, the use of non-expert intuition as an annotation strategy raises a methodological question, which is to what extent can subjective preferences reliably capture a construct that has identifiable linguistic correlates. We address this question by modeling evaluators’ textual preferences and comparing the resulting classifications with linguistic features that previous research has independently associated with professional communication. This allows us to assess not only whether professionalism can be computationally modeled, but also whether non-expert preference provides a meaningful basis for its annotation.

Our field of application is financial communication, a domain where professionalism is both expected and strategically deployed; this study in particular will handle question turns in analyst-corporation dialogical exchanges. This choice is supported by a few reasons. First and foremost, it is intrinsically a highly professional environment – which means that regardless of the evaluations that we will collect and employ, the ground truth is that **every** naturally occurring production in this field is professional. Consequently, our results cannot be exploited for nefarious purposes such as discrimination in the workplace. Secondly, there is wide availability of freely accessible transcripts of such exchanges, which makes data collection relatively easy and replicability straightforward. Third, being

the transcriptions readily accessible documents on the web, we can infer that it is plausible that they are included in the training corpus of most modern LLMs, which therefore are acquainted with the genre. LLMs will be extensively used in our study as a SotA means of comparison. Fourth, questions in this domain are constrained and formulaic, which allows for clear segmentation and recognition of patterns.

Our approach is oriented at result interpretability, methodological transparency and resource use efficiency. We probe whether an elusive concept like professionalism – which could however be theoretically modeled by a linguistic framework that builds on politeness and compliance with a number of constraints set by the activity type (Rigotti and Rocci, 2006) – is a consistent intuitive notion across non specialists, and if so, whether we are capable of modeling it through explicit parameters which allow for cost-efficient classification. For this purpose, our analysis employs concepts and tools borrowed from theoretical linguistics. To this end, we leverage on previous contributions in this domain such as the linguistic question characterization by Lucchini and D’Agostino (2023) and prior research that utilizes it such as D’Agostino et al. (2024). They already build and validate a theoretical framework, relieving us from such a burden.

This study deliberately combines synthetic and natural data to investigate professionalism perception independently from text origin. This is motivated by the acknowledgment that modern communicative exchanges – including those in professional settings – are increasingly populated by synthetic content, often emulating human productions very efficiently. While this raises both ethical and practical concerns (Cardon and Coman, 2025), it is a phenomenon that demands contextualized research.

Methodologically, we compile a dataset of financial Q&A questions (balanced between natural and synthetic), and collect multiple human ratings of perceived professionalism. Each question is also annotated with features that explicitly capture its textual characteristics. Our findings show that these features effectively describe questions in a way that enables simple classifiers to learn and reliably predict the empirical perception of professionalism – better than SotA LLMs.

We list our contributions as follows:

- a. We model readers’ textual preferences regarding professionalism, providing an empirical account of how non-expert evaluators perceive professional quality in questions within a controlled domain.
- b. We investigate whether these preferences align with linguistically motivated features of professional communication, as well as with the natural/synthetic distinction, thereby assessing the extent to which non-expert preference can serve as a meaningful basis for the annotation and computational modeling of professionalism.
- c. We show that these preferences can be modeled through a lightweight and interpretable classification approach, identifying the combination of linguistic features that best predicts perceived professionalism while preserving methodological transparency.

2 Related Work

2.1 Professionalism

Professional communication encompasses domain-specific practices through which participants pursue institutional goals and convey professional identities (Faber, 2002; Sarangi and Roberts, 1999). These practices involve communicating clear and appropriately framed information (Couture, 1992), persuading interlocutors, and maintaining working relationships (Beaufort, 2007). Their linguistic realization is shaped by institutional roles and communicative purposes (Drew and Heritage, 1992), including conventions governing how questions are formulated and interpreted (Clayman et al., 2007; Ilie, 2021).

Financial NLP has also investigated computational indicators of professional quality. Chen et al. (2021) use complementary stylistic and semantic features to distinguish professional analysts’ rationales from those of amateur investors and identify expert-like financial opinions. Chen et al. (2024) introduce professionalism-aware pre-finetuning to improve profitability ranking of investor opinions. These studies connect textual characteristics with professional expertise and downstream financial utility. Our target is different: we model *perceived professionalism*, namely recipients’ judgments of whether a question aligns with their expectations of professional communication. Rather than inferring the questioner’s actual competence or predicting investment outcomes, we investigate the interpretable

linguistic features associated with these judgments.

2.2 Financial Q&A

The Q&A sessions of earnings conference calls (ECCs) provide a setting in which financial analysts question corporate representatives about company performance and prospects. Prior research describes the genre conventions of ECCs and the formulation of questions as delicate interactional acts (Crawford Camiciottoli, 2010; de Oliveira and Pereira, 2018), while financial studies examine the informational significance of conference calls and manager–analyst exchanges (Kimbrough, 2005; Chen et al., 2018). Analysts’ questions can request information, invite explanations, or challenge corporate positions (Palmieri et al., 2015). Linguistic accounts characterize these practices through question structure, prefacing, and pragmatic functions (Lucchini and D’Agostino, 2023; D’Agostino et al., 2024). These perspectives motivate attention to *competency*, *trustworthiness*, and strategically deployed politeness, rather than surface formality alone.

Complementary NLP research addresses the generation of potential analyst questions. Juan et al. (2023) investigate multiple-question generation from presentation transcripts and demonstrate the benefits of key-point retrieval. Juan et al. (2024) develop a co-trained retriever–generator framework for earnings-call questions and include professionalism in their human evaluation, using annotators with financial expertise. We complement these generation-oriented studies by examining the linguistic correlates of perceived professionalism across natural and synthetic questions, rather than treating professionalism solely as an output-quality criterion.

2.3 Politeness

Politeness concerns the context-sensitive management of social relations and face in interaction (Brown and Levinson, 1987; Watts, 2003). In ECCs, it can support professional identity negotiation and persuasive communication (Koller and Wu, 2023; Crawford Camiciottoli, 2011). Friendly or indirect formulations may therefore serve strategic purposes rather than merely signal formality. We use *loaded politeness* to refer to this strategic dimension of professional interaction.

Human–LLM comparisons further indicate that perceived courtesy need not imply equivalent pragmatic behavior. Zhao and Hawkins (2025) find that

LLMs disproportionately rely on negative politeness strategies, even in contexts favoring rapport-building, whereas Park et al. (2025) identify difficulties in responding to relational cues in Korean conversational scenarios. These findings motivate examining how natural and synthetic questions differ in their interactional features, without equating greater courtesy with greater professionalism. We do not collect separate politeness ratings or attempt a comprehensive operationalization of strategic politeness. Instead, we examine selected interactional features in relation to perceived professionalism and return to their implications in Section 4.

2.4 Non-expert evaluation of complex phenomena

The suitability of non-expert judgments depends on the construct and task being evaluated. Snow et al. (2008) show that aggregating non-expert annotations can approximate expert labels on several NLP tasks. For subjective discourse quality, Behrendt et al. (2024) combine expert annotations of deliberative dimensions with non-experts’ perceptions to construct an interpretable quality score. This provides a methodological precedent for relating theoretically motivated descriptors to lay judgments, rather than assuming that expert and non-expert assessments are interchangeable. Audience-centered evaluation is also relevant to financial text generation. Takayanagi et al. (2025) assess generated investment analyses through their effects on amateur and professional investors’ decisions, alongside judgments of textual qualities. Their approach complements our focus on recipients’ responses, although decision influence and perceived professionalism are distinct evaluation targets.

LLMs offer another source of annotation, but their suitability also requires task-specific validation. Tseng et al. (2025) evaluate LLM annotation across finance, biomedicine, and law, finding that reasoning and inference-time techniques do not consistently improve performance. Accordingly, we separately assess LLMs as linguistic-feature annotators and as predictors of professionalism judgments. Our collection of non-expert ratings on a three-point scale is intended to capture perceived professional quality, not to establish that lay judgments can substitute for expert assessments of substantive question quality.

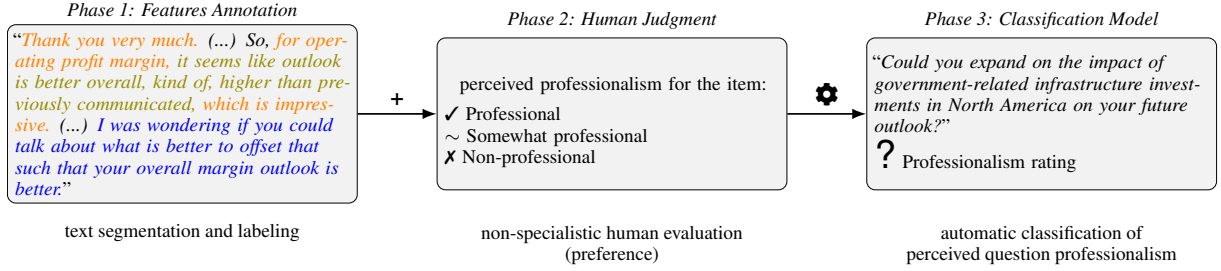


Figure 1: Three-phase architecture of the study.

3 Study

Our study encompasses three experimental phases. **Phase one** automatically annotates theoretically-motivated features and evaluates the annotation performance.

Phase two empirically collects non-expert perception of professionalism across and regardless of question origin (natural vs. synthetic).

Phase three builds a regression model of professionalism prediction based on linguistic parameters and deploys a prediction tool to classify questions with performance better than the baseline.

3.1 Phase one

3.1.1 Features

As shown by previous qualitative work (Palmieri et al., 2015; Rocci and Raimondo, 2017; D’Agostino et al., 2024; Lucchini et al., 2024) ECC questions can be effectively characterized by linguistic features that are theoretically grounded in discourse analysis, pragmatics, and question typology. The linguistic features of choice for this work were originally described by Lucchini and D’Agostino (2023). We divide them into four categories:

Discourse regulators Dialogical or meta-discursive components that organize the interactional flow of the question turn. They may acknowledge the previous speaker (“Thanks”), specify who the question is for (“Chris, this one is for you”), indicate the number or order of questions (“I have two questions”), or introduce the topic explicitly (“A quick one on margins”). They mediate the exchange in ways that are socially and structurally relevant.

Prefaces Assertions that introduce, shape, or justify the interrogative act. They may state facts (“Dividends are up”), express opinions (“It seems like guidance is too optimistic”), report others’

speech (“As you mentioned last quarter”), or include meta-commentary (“I ask this because it wasn’t clear to me”). They are often used to construct credibility and soften face-threatening acts.

Question types The syntactic formulation of an interrogative act. We consider three types: open, polar, and closed-list (multiple choice).

Request types They describe the pragmatic value of the speech act that is instantiated in the question; the underlying assumption is that every question asked in this context is a request. This level of pragmatic annotation allows us to approximate the functional load of each question.

To the best of our knowledge, no attempt was ever made to automatically identify such manually annotated features at scale. We propose to test and evaluate LLMs on this task, assuming that such models constitute SotA for linguistic applications. We compare gpt-4o, gemini-1.5-pro, gemini-2.0-flash, and claude-3.7-sonnet, tested on each annotation task (i.e., a single linguistic feature) independently. LLM calls are sent through API with a chat-completion request; hyperparameter settings are left at default value.¹

Output performance is assessed in terms of F_1 and Γ ² accuracy. Subsequent phases will build upon the automatic annotations produced by the best-performing model for each task, namely gemini-2.0-flash for discourse regulators, gpt-4o for prefaces, gemini-1.5-pro for question types, and claude-3.7 for request types. As the two metrics were often misaligned, the selection criterion was F_1 accuracy.

The prompt templates are available in Appendix A. Table 4 in Appendix C reports performance evaluation across the four models, compared against

¹Default temperature=1.0

²Method for inter-annotator agreement measure and alignment (Mathet et al., 2015); Python library pygamma-agreement

human IAA on the same task as a baseline. While results vary across tasks – with prefaces annotation performed best and request types being the worst – the human baseline consistently outperforms LLMs. Additional task-specific details on error sources and types for the best-performing model are provided in Appendix B.

These linguistic features are furthermore paired with the automatic annotation of a number of more traditional text attributes – loosely inspired by Reddy et al. (2021)’s analysis – included to enrich questions representation, with the objective of strengthening predictive performance. These traditional NLP features are: Flesch-Kincaid readability score, Dale-Chall readability score, sentence count, question count, mean question length, assertion count, mean assertion length, filler word count, interjection count, stopword count, person count (NER), type-token ratio, vocabulary entropy.

All features leveraged for classification in this study are explicitly handcrafted. Consequently, they may not fully capture the entire range of relevant descriptors for our categories and may therefore limit classification effectiveness. In the worst case, such a choice could introduce bias into the analysis. We argue, however, that the design of our approach mitigates this risk: handcrafting features does not in itself entail subjective judgments, and the soundness of the selected features is supported both by previous domain-specific research and by algorithmic validation (see Section 3.3.1). We acknowledge the related concern that such features may not be sufficiently expressive to capture the full complexity of our categories. At the same time, we believe that this approach offers a key advantage for the present work, namely, the full interpretability of the resulting analyses, which is one of its main objectives.

Features and question origin Feature-based characterization of question origin exceeds the scope of this paper, but it is worth mentioning that the features selected in this phase proficiently discern between natural and synthetic questions. Spearman correlation testing shows that most features align with natural texts, whereas synthetic questions are decidedly associated with high readability and variability (type-token ratio); see Appendix D for further details. This serves as an additional reason supporting the scientific interest on whether they also efficiently describe non-expert preference alignment.

3.2 Phase two

Our original dataset comprises 500 question units (250 natural, 250 synthetic). The prompt template for synthetic questions generation is available in Appendix E. Further information about dataset building can be found in Appendix F.

We supplemented our original dataset with the dataset introduced by Juan et al. (2024)³, which comprises 250 question items. Although the provenance of individual questions in this dataset could not be independently established, we included it based on the authors’ confirmation that datapoints are balanced between natural and synthetic items, consistently with our own data.

The full dataset thus encompasses 750 items, balanced across question origin. The relatively small size of the dataset compared to traditional NLP resources of akin fields (e.g., other benchmarking datasets available for detection tasks of emotions (Buechel and Hahn, 2017) or trustworthiness (Huang et al., 2024), or evaluation tasks of veracity (Gupta et al., 2022)) is not troubling: considering the initial costs of both manual and synthetic annotations, it seems sensible that only a limited quantity of resources is allocated for an exploratory study – any outcome of which is novel regardless. A more extensive dataset may be collected and annotated for future work, based on (and, possibly, adapted to) the results of the current study.

Each question in Juan et al. (2024)’s work was rated by three independent annotators on a 3-point scale for perceived professionalism. For consistency, we crowdsourced⁴ professionalism annotation for our original dataset following a similar method. Each respondent rated 10 items and each item was annotated by 5 different raters (refer to Appendix G for participant instructions). Eligible raters were invited to participate in the study according to three criteria: (1) they were native speakers of English, (2) they were UK or USA nationals, (3) they held a university degree. No constraints were set on the respondent’s field of specialization.

The multiple professionalism evaluations per item were tested for internal consistency by calculating Cronbach’s Alpha, which proved acceptable ($\alpha = 0.70$). This licensed the use of their

³Juan et al. (2024) shared their dataset upon request and we have no reasonable grounds for doubting their academic integrity. Thus, we trust their word on dataset quality and the methodology with which they built it.

⁴<https://www.prolific.com/>

mean as the reference value for their perceived professionalism, rounded to the closest integer. The resulting professionalism score is not uniformly distributed and is arranged as follows: score 1 (non-professional): 5 items; score 2 (somewhat professional): 582 items; score 3 (professional): 163 items.

These results, and particularly the marginality of the “non-professional” class, support the ground truths stated earlier: (a) all natural questions are professional by definition, and (b) LLMs are rather good at mimicking a style that can be evaluated to some extent professional by non-experts, also due to their highly-polite turn formulations.

To conclude this phase, all items were also automatically annotated for the features presented in phase one.

Relationship between professionalism and question origin A point-biserial correlation testing immediately shows that perceived professionalism correlates negatively with question authenticity: $r = -0.31, p < 0.05$. This means, it is on average more plausible that a synthetic question is evaluated more professional than a natural one, with statistical significance. This is in line with previous research (Coman and Cardon, 2024), according to which the characteristics of synthetically generated text are generally perceived as reassuring and soothing – which leads to a more positive assessment of it, apparently in term of professionalism too. Consequently, question origin can be practically interpreted as a confounding factor, which is however factored in the experimental design.

3.3 Phase three

In phase three we pursue two main objectives.

First, we investigate the optimal combination of features from phase one that describe our professionalism classes. To reach this objective, we run four different feature sets through a regression model (Ordinary Least Squares, OLS), which is a statistical technique used to examine the relationship between a dependent variable and one or more independent variables. We consider professionalism the dependent variable – that is, we investigate whether the model is fit to represent it – while our features constitute the independent variables. Count variables are normalized by text length; continuous variables are standardized.

Still, in phase two we observed that only five items received score “non-professional”. So few

Features	R^2	AIC	BIC	Omnibus
linguistic	0.298	770.7	890.8	85.268
NLP	0.351	685.3	745.4	86.229
combined	0.379	704.2	884.4	72.969
combined (optim.)	<u>0.378</u>	<u>703.0</u>	878.5	73.471

Table 1: OLS regression results, binary classification. The best performing results are in bold, second best underlined.

data points are hardly enough for generalization. At the same time, we suspect that, although marginal, such a class is significantly distinct from the “somewhat professional” one – and, therefore, to arbitrarily assimilate the two may skew results. Accordingly, this first sub-phase compares the original three-way classification with a binary classification, which serves the purpose of verifying the actual independence of the non-professional class. For the binary classification, the “non-professional” (as opposed to “professional”) class condenses the *non-professional* and *somewhat professional* categories from the survey. We expect only one classification heuristic among the two to prove valid.

Secondly, we run an automatic classification experiment across machine learning models. Here the objective is to identify the optimal classification method that allows to distinguish between the different categories of *perceived professionalism*. The feature sets we employ in this second sub-phase are the two best feature combinations of the most fitting classification heuristic. Feature combination grade is evaluated on the basis of highest R^2 and lowest AIC scores obtained from the prior OLS regression testing.

3.3.1 Features optimization

Regression is run on four sets of features: (a) the linguistic features previously introduced, (b) NLP traditional features, (c) the combination of the previous two, and (d) an optimized subset of the “combination” set. Such optimization is achieved running recursive feature elimination (RFE) with cross-validation (CV). The number of features selected is tuned automatically by fitting an RFE selector on the different cross-validation splits; estimation is performed by an SVR.

3.3.2 Regression results

Results of the regression tests are reported in Table 1 and Table 2 for the binary and three-way classification tasks respectively. The detailed regression report for the two best-performing configurations

Features	R^2	AIC	BIC	Omnibus
linguistic	0.863	1935.	2055.	72.504
NLP	0.983	364.1	424.1	8.037
combined	0.983	389.6	569.8	6.414
combined (optim.)	0.983	<u>387.9</u>	563.4	6.242

Table 2: OLS regression results, three-way classification. The best performing results are in bold, second best underlined.

is available in Appendix H.

Binary classification The most immediate consideration is that binary classification struggles to keep pace with the full three-classes alternative, thus hinting that the “non-professional” category may in fact possess distinct characteristics from the “somewhat professional” class to which it was aggregated. Therefore, this classification heuristic is discarded, as it misrepresents the data.

Three-way classification Considering R^2 and AIC as indicators, it appears that NLP features alone already accurately describe the textual differences between the three categories. Second-best performing is the optimized combination of linguistic and NLP features, thus allegedly showing that the injection of novel linguistic features only lowers classification performance. This is to some extent consistent across classification heuristics: also for binary classification, the NLP set hits lowest AIC. Also in light of the high R^2 results obtained, we reckon that this may be an effect of model overfitting – particularly, on the measly represented “non-professional” category.

Consequently, we acknowledge that there are too few examples to model the “non-professional” class reliably⁵, which will therefore be removed from the later classification experiment. The classification experiment that follows thus only keeps the remaining “somewhat professional” and “professional” classes of the original three-way classification. Ostensibly, this is the best strategy available to not compromise experimental design.

To verify the soundness of our preliminary results, the classification experiment compares the two best-performing feature sets (NLP only vs combined

optimized) across machine learning models.

3.3.3 Automatic classification

The baseline for our classification experiment is trained over raw question texts. It encompasses (a) an SVM, and (b) a 20-shot instruction-tuned LLM. The LLM is a GPT model, since the GPT family is generally understood as one of the best performing architectures for such tasks, which could be accessed through API, and GPT-5.4 was the latest model available on the market at the time of the experiment.⁶ In-context learning features 20 examples, as we assessed that iterations with 40 and 60 examples exhibited stable or decreased performance; the examples were equally distributed between professional and non-professional instances, averaged by question origin. Furthermore, prompt tuning allowed us to retain the prompt formulation that returned best outcomes.

We compare the baseline against four classification approaches: (a) Random Forest classifier trained over features only, (b) Random Forest classifier trained over DistilBERT text embeddings + features, (c) DistilBERT fine-tuned over text embeddings + features and (d) Logistic Regression classifier trained over features. The dataset is split into the same train and test sets with a 80:20 ratio for every classification approach.⁷ Results are reported in Table 3.

Results show that basic text-based classification (baseline) achieves fair accuracy, thus demonstrating that perceived professionalism truly holds onto some textual characteristic of the question. For both feature sets explored, the best performing approach appears to be the one labeled “DistilBERT + RF”, which is the combination of DistilBERT-based embeddings, features, and a Random Forest classifier. In line with our previous OLS regression results, the performance gap between the two feature sets (NLP features only vs an optimized combination of linguistic and NLP features) is negligible; however, in contrast with what seen before, it is the combined set that unexpectedly achieves a slightly higher F_1 score.

It is worth noting that the SotA GPT model performs quite poorly. The issue does not seem to be a general model under-performance in linguistic classification tasks: upon further inspec-

⁵We remind the reader that this striking class imbalance is a byproduct of the second stage of this study (categorization based on user preference), whereas items were originally perfectly balanced according to objective criteria. The distribution asymmetry simply cannot be corrected, as it reflects a real-world tendency, and we verified that forcefully augmenting the minority class causes more harm than good.

⁶As of May 2026.

⁷Exception: 20-shot in-context learning for the LLM is tuned, by definition, on 20 examples – but it is tested on the same 100 examples of the 20% test-set as the other models.

	<i>Model</i>	<i>Accuracy</i>	<i>Macro-P</i>	<i>Macro-R</i>	<i>Macro-F1</i>
Baselines					
<i>deterministic baseline</i>	SVM	0.77	0.58	0.56	0.57
<i>stochastic baseline</i>	GPT-5.4 (20-shot)	0.42	0.54	0.53	0.41
NLP features only					
	Random Forest	0.78	0.60	0.58	0.59
	DistilBERT + RF	0.82	<u>0.68</u>	0.61	<u>0.63</u>
	DistilBERT	<u>0.80</u>	0.45	0.46	0.45
	LogRegression	0.82	0.69	0.61	<u>0.63</u>
Combined features (optimized)					
	Random Forest	<u>0.80</u>	0.65	0.62	<u>0.63</u>
	DistilBERT + RF	0.82	<u>0.68</u>	<u>0.63</u>	0.64
	DistilBERT	0.78	<u>0.62</u>	0.61	0.62
	LogRegression	0.63	0.61	0.68	0.58

Table 3: Comparison of models, across features combinations, on the professionalism classification task for the two representative classes of the three-way heuristic (best results in bold, second best underlined). The two models on top are treated as text-based baselines – deterministic and stochastic respectively.

tion, GPT-5.4 perfectly classifies question origin ($F_1 = 1.00$) when using the same prompt, lexically adapted to the task, and an equal number of examples across the two classes.⁸ It therefore may be the case that the features characterizing perceived professionalism – or the combination thereof – elude the LLM’s scrutiny. However, it seems that the LLM does not classify question origin instead, as no statistically significant evidence of an association can be found between the two variables ($\chi^2(1, N = 100) = 1.606, p = 0.205$).

4 Discussion

Results highlight the following main contributions.

Due to a carefully drafted feature selection and the computational optimization of their combination, we determined an optimal set of features that describe better than the baseline perceived professionalism evaluation. Such an evaluation is intended as a proxy for the investigation of preference alignment on professional questions as perceived by non-experts. The goal is to verify whether professional question quality can be reliably assessed by non-experts and/or LLMs; arguably not.

The relatively high degree of accuracy already yielded by the deterministic baseline (SVM) trained on the sole question text supports the finding that traditional NLP features alone effectively discriminate between the perceived more or less professional tone of a question. At the same time, this also confirms that non-expert preference lies on surface traits. Accordingly, non-experts also tend to prefer synthetic questions – that is, those that most distinctively present a simple yet clear and direct

surface structure – which is a finding consistent with the literature.

Remarkably, domain-specific nuances of politeness, such as friendliness and chit-chatting (already discussed in Section 2.3), are neither captured by LLMs nor appreciated by evaluators. In fact, discourse regulators such as “greeting” or “outside comment” are completely absent from synthetic questions; Table 5 in Appendix D reports the correlation strength of these features with natural questions. At the same time, evaluators systematically prefer turns exhibiting little to no use of such features.

The results are consistent with the hypothesis that non-expert judgments rely more strongly on surface characteristics than on domain-specific features identified in previous research. Our findings suggest that expertise cannot be replaced by candid evaluation, which entirely lacks domain knowledge and expertise.

Conversely, the sheer effortlessness with which question origin is automatically categorized signals that significant work is still required towards optimization of questions generation – particularly for high-stakes professional domains. While current results are not ripe enough to provide a comprehensive set of features and guidelines that, if applied, would grant the generation of items virtually indiscernible from natural occurrences, we argue that this study provides a clear starting point in the direction of effective modeling through suitable (and interpretable) feature design.

⁸This result also retroactively supports prompt quality.

5 Conclusions and Future Work

In this work, we combine empirical insights from social science research with theoretical considerations of linguistic features to develop an interpretable NLP approach to the study of perceived professionalism in expert communication.⁹ The resulting best-performing model effectively discriminates between more and less professional questions, providing a first computational, interpretable formulation of non-experts' evaluation of professional quality. By examining non-experts' preferences in a professional domain, our study also provides insights into the validity of non-expert evaluation as a source of evidence for linguistic research and the potential value of professional evaluation in domain-specific applications. Finally, it emerges that synthetic questions score highly on perceived professionalism, but at the same time, they still substantially differ from natural questions in domain-specific interactional patterns.

A first contribution of this work is the annotated dataset itself, which provides a preliminary benchmark for the study of perceived professionalism in questions, together with a methodology for extracting transparent linguistic features associated with this construct.

Moreover, despite the limited size of the dataset and the constraints imposed by the use of hand-crafted features in the classification experiment, our findings already point to several directions for future research.

NLP research Our results show that effective classification can be achieved with relatively simple architectures when these are grounded in theoretically motivated linguistic features. This finding supports a renewed consideration of explicit linguistic knowledge in NLP research, particularly in language generation. A natural next step is to investigate whether the identified features can be used to optimize LLM-generated questions with respect to both perceived and objective professionalism.

Professional practice Beyond academic research, effective systems for generating synthetic professional questions could have immediate applications in professional education and training. In

particular, they could support the training of early-career managers by providing exposure to well-formed, high-quality questions and thereby helping develop effective communicative responses. Conversely, synthetically generated professional answers could provide a benchmark against which the effectiveness of naturally occurring answers to the same questions could be evaluated.

Limitations

The primary limitation of this study is the relatively limited size of the dataset, which was also constrained by the costs associated with data collection and annotation. This limitation is partly offset by the richness and complexity of the dataset. Each item consists of a complete question turn, encompassing multiple topical units, each of which exhibits a distinct linguistic and strategic structure. Each element of these structures is systematically annotated according to the procedure described in phase one. The resulting dataset therefore provides a rich and densely annotated resource for NLP research, which can be extended by future studies.

The class imbalance observed in phase three, as discussed in the paper, is a structural consequence of phase two and reflects a tendency observed in the collected data. Rebalancing the dataset by returning to the original pool and adding items to the underrepresented class would have altered this distribution artificially and would have been prohibitively expensive. We also considered augmenting the minority class, but its degree of underrepresentation resulted in model overfitting when this approach was tested. We therefore retained the original distribution, which we consider the most appropriate solution given the constraints of the study.

Finally, responses to professional and less professional questions were deliberately left outside the scope of the present study. Their evaluation would require a different methodological framework, owing to their considerably greater length and complexity, which makes it difficult to obtain reliable judgments through an online evaluation platform. The evaluation of such responses remains largely unexplored and represents a promising direction for future research.

Ethics statement

The human judgments collected for this paper were gathered through Prolific. We recruited external

⁹The datasets presented in this paper, including professionalism judgments and automatic annotation of features, are available from the dedicated GitHub repository: https://github.com/dagosgi/FinNLP2026-perceived_professionalism

annotators but did not process any of their personal data; participant selection criteria were set as initial parameters (screeners), processed by the annotation platform entirely. Participants were aware that they could withdraw from the study any point and their response would not be recorded. The task only consisted in the selection of the perceived professionalism value (3-point scale) for each of the 10 items proposed. According to the participation guidelines, each respondent who completed the study received compensation, in the amount of £6/hour for synthetic questions rating and £7.50/hour for natural questions rating. Those numbers were estimated a fair compensation for the task by the platform automatic assessment. Ultimately, we do not identify ethical concerns associated with our study.

Acknowledgments

This paper is based on results obtained from AIST policy-based budget project “R&D on Generative AI Foundation Models for the Physical Domain.” The work of Giulia D’Agostino was supported in part by the Swiss National Science Foundation under the project “Mining argumentative patterns in context. A large scale corpus study of Earnings Conference Calls of listed companies” (grant n. 200857) and in part by the internship programme of the National Institute of Advanced Industrial Science and Technology (AIST) in Japan.

References

- Anne Beaufort. 2007. Writing in the Professions. In *Handbook of Research on Writing*. Routledge.
- Maike Behrendt, Stefan Sylvius Wagner, Marc Ziegele, Lena Wilms, Anke Stoll, Dominique Heinbach, and Stefan Harmeling. 2024. [AQuA – Combining Experts’ and Non-Experts’ Views To Assess Deliberation Quality in Online Discussions Using LLMs](#). In *Proceedings of the First Workshop on Language-driven Deliberation Technology (DELITE) @ LREC-COLING 2024*, pages 1–12, Torino, Italia. ELRA and ICCL.
- Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*, 1 edition. Cambridge University Press.
- Sven Buechel and Udo Hahn. 2017. [EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.
- Laura Yvonne Bulk, Donna Drynan, Sue Murphy, Patricia Gerber, Roberta Bezati, Sacha Trivett, and Tal Jarus. 2019. [Patient perspectives: Four pillars of professionalism](#). *Patient Experience Journal*, 6(3):74–81.
- Peter W. Cardon and Anthony W. Coman. 2025. [Professionalism and Trustworthiness in AI-Assisted Workplace Writing: The Benefits and Drawbacks of Writing With AI](#). *International Journal of Business Communication*, page 23294884251350599.
- Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. [Evaluating the rationales of amateur investors](#). In *Proceedings of the Web Conference 2021, WWW ’21*, pages 3987–3998, New York, NY, USA. Association for Computing Machinery.
- Chung-Chi Chen, Hiroya Takamura, Ichiro Kobayashi, and Yusuke Miyao. 2024. [Professionalism-aware pre-finetuning for profitability ranking](#). In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM ’24*, page 3674–3678, New York, NY, USA. Association for Computing Machinery.
- Jason V. Chen, Venky Nagar, and Jordan Schoenfeld. 2018. [Manager-analyst conversations in earnings conference calls](#). *Review of Accounting Studies*, 23(4):1315–1354.
- 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.
- Anthony W. Coman and Peter Cardon. 2024. [Perceptions of Professionalism and Authenticity in AI-Assisted Writing](#). *Business and Professional Communication Quarterly*, page 23294906241233224.
- Barbara Couture. 1992. [Categorizing Professional Discourse: Engineering, Administrative, and Technical/Professional Writing](#). *Journal of Business and Technical Communication*, 6(1):5–37.
- Belinda Crawford Camiciottoli. 2010. [Earnings calls: Exploring an emerging financial reporting genre](#). *Discourse & Communication*, 4(4):343–359.
- Belinda Crawford Camiciottoli. 2011. [Ethics and ethos in financial reporting: Analyzing persuasive language in earnings calls](#). *Business Communication Quarterly*, 74(3):298–312.
- 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.
- Brenton Faber. 2002. Professional identities: What is professional about professional communication? *Journal of Business and Technical Communication*, 16(3):306–337.
- Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. *DialFact: A benchmark for fact-checking in dialogue*. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801, Dublin, Ireland. Association for Computational Linguistics.
- Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Hanchi Sun, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, and 52 others. 2024. *Trustlm: Trustworthiness in large language models*. In *Forty-first International Conference on Machine Learning*.
- 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. 2023. *Generating Multiple Questions from Presentation Transcripts: A Pilot Study on Earnings Conference Calls*. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 449–454, Prague, Czechia. Association for Computational Linguistics.
- Yining Juan, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2024. *Co-trained retriever-generator framework for question generation in earnings calls*. Preprint, arXiv:2409.18677.
- Michael D. Kimbrough. 2005. *The Effect of Conference Calls on Analyst and Market Underreaction to Earnings Announcements*. *The Accounting Review*, 80(1):189–219.
- Veronika Koller and Xiaoxi Wu. 2023. *Analysts’ identity negotiations and politeness behaviour in earnings calls of us firms with extreme earnings changes*. *Corporate Communications: An International Journal*, 28(5):769–787.
- So Ran Kwon, Udochukwu Oyoyo, Connor Hendry, Nouf Maayta, Lourd Nafea, and Vania Aranda. 2025. *Scrub Color Affects Perceived Professionalism of Dental Hygienists: A Cross-sectional Study*. *The Journal of Contemporary Dental Practice*, 26(3):223–228.
- Costanza Lucchini and Giulia D’Agostino. 2023. *Good answers, better questions. Building an annotation scheme for financial dialogues*. Technical report.
- Costanza Lucchini, Andrea Rocci, and Olena Yaskorska-Shah. 2024. *Comparing prefaced questions across activity types : journalists and financial analysts as argumentative questioners*. *De l’argumentativité à l’argumentation*. In press.
- Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. *The Unified and Holistic Method Gamma (Γ) for Inter-Annotator Agreement Measure and Alignment*. *Computational Linguistics*, 41(3):437–479.
- Katrin Ortmann. 2022. *Fine-Grained Error Analysis and Fair Evaluation of Labeled Spans*. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1400–1407, Marseille, France. European Language Resources Association.
- Rudi Palmieri, Andrea Rocci, and Nadzeya Kudraut-sava. 2015. *Argumentation in earnings conference calls. Corporate standpoints and analysts’ challenges*. *Studies in communication sciences*, 15, 2015(1):120–132.
- Seoyoon Park, Jaehee Kim, and Hansaem Kim. 2025. *Too polite to be human: Evaluating LLM empathy in Korean conversations via a DCT-based framework*. In *Proceedings of the Third Workshop on Social Influence in Conversations (SICon 2025)*, pages 76–89, Vienna, Austria. Association for Computational Linguistics.
- Sravana Reddy, Mariya Lazarova, Yongze Yu, and Rosie Jones. 2021. *Modeling language usage and listener engagement in podcasts*. 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 632–643, Online. Association for Computational Linguistics.
- E. Rigotti and A. Rocci. 2006. *Towards a definition of communication context. foundations of an interdisciplinary approach to communication*. *Studies in Communication Sciences*, 6(2):155–180.
- Andrea Rocci and Carlo Raimondo. 2017. *Conference calls: A communication perspective*. In Alexander V. Laskin, editor, *The Handbook of Financial Communication and Investor Relations*, pages 293–308. John Wiley & Sons, New York, NY.
- Srikant Sarangi and Celia Roberts. 1999. *Talk, Work and Institutional Order: Discourse in Medical, Mediation and Management Settings*. Mouton de Gruyter, Berlin.

- Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. [Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks](#). In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.
- Takehiro Takayanagi, Hiroya Takamura, Kiyoshi Izumi, and Chung-Chi Chen. 2025. [Can GPT-4 sway experts' investment decisions?](#) In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 374–383, Albuquerque, New Mexico. Association for Computational Linguistics.
- Yu-Min Tseng, Wei-Lin Chen, Chung-Chi Chen, and Hsin-Hsi Chen. 2025. [Evaluating large language models as expert annotators](#). In *Second Conference on Language Modeling*.
- Richard J. Watts. 2003. *Politeness*, 1 edition. Cambridge University Press.
- Haoran Zhao and Robert D. Hawkins. 2025. [Comparing human and LLM politeness strategies in free production](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16188–16216, Suzhou, China. Association for Computational Linguistics.

A Prompts for linguistic features annotation

Discourse regulators You are an expert in extracting and classifying discourse regulators from text.

We define discourse regulators as a class of turn fillers the minimal length of which is the clause. Discourse regulators do not comprise questions. Additionally, discourse regulators are not prefaces to questions; prefaces are portions of text that introduce content to the conversation, whether facts, opinions, or reported speech. You only analyze discourse regulators found in question turns.

Discourse regulators belong to the following types:

- recipient: indication of whom the question is addressed to
- theme: the topic about which the question is going to be asked
- enumeration: indication of how many questions the turn will (or did) contain
- counting: progressive indication of the number/order of the question that is being asked
- greeting: greeting towards the interlocutor(s)
- acknowledgment: acknowledgment for turn taking and interlocutor's attention
- agreement: agreement to a previous turn (typically, in response to an answer)
- inside comment: comment about a intra- or inter-turn fact or opinion, or about a previous turn
- outside comment: comment about a fact or opinion outside the conversation

If more than one type seems to be present in one clause, the labeling favors the type higher in the list.

You are requested to extract and classify discourse regulators found in question turns. Your output must be a JSON object of the format: 'discourse regulator 1': 'value 1', 'discourse regulator 2': 'value 2', ...

Examples:

INPUT: 'Thank you. That's super helpful. And then just a quick one on POS. I know you mentioned POS down for Q1, which is not surprising, given that Easter shift. I was curious if you could share, if you have, you know, year-to-date Easter adjusted numbers that include the first two weeks of POS in April. And then while we are at it, do you have a week of inventory at retail at quarter end handy?'

OUTPUT: 'Thank you': 'acknowledgment', 'then just a quick one on POS': 'theme', 'then while we are at it': 'counting'

INPUT: 'Great. And then just a quick follow-up for Darren, if possible.'

OUTPUT: 'Great': 'acknowledgment', 'just a quick follow-up for Darren': 'recipient'

INPUT: 'Darren, so congratulations on The Rookie getting renewed for a fifth season. That should give you close to 100 episodes at the end of next year, which should put us in a good position to sell into the syndicated – syndication market for TV. I know you're on Hulu, but wondering if you could talk about maybe some off-network syndication, how that market is shaping up, and where you are with that process.'

OUTPUT: 'Darren': 'recipient', 'congratulations on The Rookie getting renewed for a fifth season': 'outside comment'

INPUT: 'Thanks. Good morning, everyone. We have two questions. Chris, the first one is for you. It's just on helping us dimensionalize the Wizards of the Coast business between MAGIC and D&D. I think you've said in the past that you expect incremental growth to be coming from D&D over time. So share with us a little bit about the size of that business relative to MAGIC and what your expectations are. And then if you could just remind us what percentage of the Wizards business is domestic. And is there a big international opportunity? And then just a quick clarification for you, Deb, on the program amortization. Can you help us think through back half Q3 versus Q4? We just want to make sure we're modeling that line correctly going forward.'

OUTPUT: ‘Thanks’:‘acknowledgment’, ‘Good morning, everyone’:‘greeting’, ‘We have two questions’:‘enumeration’, ‘Chris, the first one is for you’:‘recipient’, ‘It’s just on helping us dimensionalize the Wizards of the Coast business between MAGIC and D&D’:‘theme’, ‘And then just a quick clarification for you, Deb’:‘recipient’, ‘on the program amortization’:‘theme’

INPUT: ‘Thanks. And then my follow-up, as it relates to Europe, I just want to clarify, you mentioned the potential risk of approximately 100 million from Russia. Are you suggesting that you could potentially ship to Russia later this year?’

OUTPUT: ‘Thanks’:‘acknowledgment’, ‘then my follow-up’:‘counting’, ‘it relates to Europe’:‘theme’

INPUT: ‘Thank you. That’s super helpful. And then just a quick one on POS. I know you mentioned POS down for Q1, which is not surprising, given that Easter shift. I was curious if you could share, if you have, you know, year-to-date Easter adjusted numbers that include the first two weeks of POS in April.’

OUTPUT: ‘Thank you’:‘acknowledgment’, ‘That’s super helpful’:‘inside comment’, ‘And then just a quick one on POS’:‘theme’

INPUT: ‘OK. Thank you.’

OUTPUT: ‘OK’:‘agreement’, ‘Thank you’:‘acknowledgment’

INPUT: ‘OK, thanks. Hey guys, good morning. I just want to go back to the margin question. Maybe looking ahead to next year, can you comment on your expectations there? You recently suggested 16%-plus was the goal. Given that you’re targeting 16% this year, is there any change to your view for next year? And then I have a follow-up.’

OUTPUT: ‘OK, thanks’:‘acknowledgment’, ‘Hey guys, good morning’:‘greeting’, ‘I just want to go back to the margin question’:‘theme’, ‘then I have a follow-up’:‘enumeration’

INPUT: ‘Got it. OK. Thanks.’

OUTPUT: ‘Got it’:‘agreement’, ‘OK’:‘agreement’, ‘Thanks’:‘acknowledgment’

Following, the text you need to analyze. INPUT: *text*.

Return ****only**** the JSON output.

Prefaces You are an expert in extracting and classifying prefaces from text.

We define a preface as a statement that accompanies a question. It can serve various purposes, such as setting the scene, reporting external facts, justifying the question, etc. Prefaces DO NOT set the topic of the question: a sentence such as, for instance, “Deb, the question is on operating margins.” IS NOT a preface. The length of the single preface spans from one clause to one sentence.

There are four classes of prefaces:

- fact: the content of the preface is presented as some factual information
- reported speech: the preface reports previous content, whether generated previously by the interlocutor or from some external source, including evaluations/opinion from third parties. It does not include the indication of topics about which people may have talked about (only contents: e.g., “I just wanted to go back to Bob’s question on margins” IS NOT reported speech – nor a preface at all), nor the acknowledgment of something that was previously done.
- opinion: the content of the preface is presented as some opinion of the speaker (an opinion could be an evaluation, an interpretation, an inference, a speculation, etc.).
- meta-discourse: the preface explicitly states the reason why asking the question was necessary.

You are requested to classify prefaces found in question turns. Your output must be a JSON object of the format: ‘preface 1’:‘value 1’, ‘preface 2’:‘value 2’, ...

Examples:

INPUT: ‘If I may just ask, but do you feel that – I mean, listening to you again, and maybe I’m interpreting it wrongly, and that’s why I’m asking the question. It sounds like you’re on the right path strategically,

i.e., top-down, you feel that you're on the right path based on the statement that you make so far. So this sounds to me like the strategic review is more an execution rather than a material top-down change in the way the business looks like, and it comes back to the how and the what. But that's kind of my interpretation from the call today.'

OUTPUT: 'listening to you again, and maybe I'm interpreting it wrongly, and that's why I'm asking the question.': 'meta-discourse', 'It sounds like you're on the right path strategically, i.e., top-down, you feel that you're on the right path based on the statement that you make so far': 'fact', 'So this sounds to me like the strategic review is more an execution rather than a material top-down change in the way the business looks like, and it comes back to the how and the what.': 'opinion'

INPUT: 'And then, getting back to the unit economics on the homes business, so it sounds like that you want to run this business with those, I forget what you call them, the railguards, whatever it is, plus 200 minus 200. And the strategy, Rich, you know, I think is you extract economics from this, then to take those economics and bring them down to the guardrail. So that long-term, you can bring down commissions, and so you can expand TAMs and just feed this beast. So that's my understanding of the flywheel. Now there are scale advantages as you do more and more of this, where did the scale advantages most show up in terms of bringing those costs down as a percentage of revenue? And, so, maybe that's really the question I'm asking here is, as you get more and more experience and more and more scale in homes business, where you think – where are you seeing and where you think the most leverage is going to come from in the expense line items that's going to allow you to, you know, get past more of those savings back to the consumer and feed that flywheel. Thanks.'

OUTPUT: 'it sounds like that you want to run this business with those, I forget what you call them, the railguards, whatever it is, plus 200 minus 200.': 'fact', 'And the strategy, Rich, you know, I think is you extract economics from this, then to take those economics and bring them down to the guardrail.': 'opinion', 'So that long-term, you can bring down commissions, and so you can expand TAMs and just feed this beast.': 'fact', 'Now there are scale advantages as you do more and more of this': 'fact'

INPUT: 'My second question is on what's left of your fixed income and equities business post the SPG divestment. What is the plan for that business overall? How much is expected to be moved in with CSFB versus how much of that market's business is to be retained by the broader CS Group? And I'd just like an idea of any thoughts on what the revenue contribution of that market's franchise could be. You've obviously given the \$2.5 billion number for CSB, but presumably, that's mainly around origination and advisory retail. Any thoughts on the markets franchise going forward?'

OUTPUT: 'You've obviously given the \$2.5 billion number for CSB': 'fact', 'presumably, that's mainly around origination and advisory retail.': 'opinion'

INPUT: 'The first one's on listings count. So you said you have similar levels to last quarter around 5.6 million. But with this hosting campaign that you guys are running, is there like a timeline that we should expect to drive that number up? Or, you know, you mentioned in the past that a quarter of guests eventually become hosts. When should we expect that number to increase?'

OUTPUT: 'you said you have similar levels to last quarter around 5.6 million.': 'reported speech', 'you mentioned in the past that a quarter of guests eventually become hosts': 'reported speech'

Following, the text you need to analyze. INPUT: *text*.

Return ****only**** the JSON output.

Question types You are an expert in extracting questions and classifying question types from text.

We define a question type as the syntactic form that a question has. Of course, this parameter can only be applied to questions.

There are three classes of question types:

- open (also known as wh-questions): there is a gap in the knowledge that needs to be filled, and this is signaled by the presence of an interrogative pronoun
- yes/no (also known as polar questions): the question is assertive and the only acceptable response values are "yes" (acceptance of the question content) or "no" (rejection of the question content)

- closed-list (also known as multiple-choice questions): two or more options are presented and one or more could be picked as a response; The last option could be an open question, that however is bound to the list-like structure of the closed-list question. Be aware that the options may be presented as consecutive sentences and not as clauses.

You are requested to classify question types found in question turns. Your output must be a JSON object of the format: 'question 1': 'value 1', 'question 2': 'value 2', ...

Examples:

INPUT: 'I have a couple of questions on investments and spending. I guess first off, what gives you the confidence that you can moderate sales and marketing spend in the second half without impacting bookings or listing levels? Is that still the plan?'

OUTPUT: 'what gives you the confidence that you can moderate sales and marketing spend in the second half without impacting bookings or listing levels ': 'open', 'Is that still the plan?': 'yes/no'

INPUT: 'And then in terms of product development, obviously, Brian, you'll be making some announcements soon. But how are you thinking about the pace of R&D or product development spend, now that business appears to be returning to normal, or whatever normal means now? Is it perhaps time to look at reaccelerating hiring? Or what are your thoughts on that?'

OUTPUT: 'how are you thinking about the pace of R&D or product development spend, now that business appears to be returning to normal, or whatever normal means now?': 'open', 'Is it perhaps time to look at reaccelerating hiring? Or what are your thoughts on that?': 'closed-list'

INPUT: 'Great, and then just one follow-up on domestic versus international. So I'm not sure if you can see this in your data yet, but just curious to hear if you're seeing some substitution in terms of domestic bookings versus international as the vaccine rates increase? And do you think the domestic rates, you know, will remain above 2019 levels even after borders open up next year?'

OUTPUT: 'just curious to hear if you're seeing some substitution in terms of domestic bookings versus international as the vaccine rates increase?': 'yes/no', 'And do you think the domestic rates, you know, will remain above 2019 levels even after borders open up next year?': 'yes/no'

INPUT: 'Brian, maybe give us your thoughts about the reopening of urban travel in terms of timing. How are you thinking about it? Is that something that you think might happen in the back half of this year? Or do you think that's further out?'

OUTPUT: 'How are you thinking about it?': 'open', 'Is that something that you think might happen in the back half of this year? Or do you think that's further out?': 'closed-list'

INPUT: 'So Brian, as a follow-up to the product announcement that you had. So, you know, I know it's probably really early, but, you know, anything you can share in terms of, you know, what their response has been? Are they delighted with what you're showing them or do you, you know, still think you have some tinkering that you need to do? Thanks.'

OUTPUT: 'anything you can share in terms of, you know, what their response has been? ': 'yes/no', 'Are they delighted with what you're showing them or do you, you know, still think you have some tinkering that you need to do?': 'closed-list'

Following, the text you need to analyze. INPUT: *text*.

Return ****only**** the JSON output.

Request types You are an expert in extracting questions and classifying request types from text.

In our texts, all questions are considered requests. The request type indicates what the question asks for. Requests belong to the following types:

- clarification
- commitment
- confirmation

- data
- elaboration
- explanation
- justification
- opinion

You are requested to extract and classify requests found in question turns. Your output must be a JSON object of the format: 'request 1': 'value 1', 'request 2': 'value 2', ...

Examples:

INPUT: 'And then similarly on the guest's side, how do you think this pandemic is going to increase willingness to use alternative accommodations? Are you seeing some positive signs there? Thank you.'

OUTPUT: 'how do you think this pandemic is going to increase willingness to use alternative accommodations?': 'opinion', 'Are you seeing some positive signs there?': 'data'

INPUT: 'Just on the supply side, will you address the issue of whether increasing local restrictions is a factor that's limiting your supply in – you know, especially in some of the larger cities? And I realize that travel hasn't really recovered there fully. But just address the issue of whether how much of a constraint new local regulations on rentals, etc., is on your ability to expand supply? Thanks a lot.'

OUTPUT: 'will you address the issue of whether increasing local restrictions is a factor that's limiting your supply in – you know, especially in some of the larger cities?': 'commitment', 'But just address the issue of whether how much of a constraint new local regulations on rentals, etc., is on your ability to expand supply?': 'elaboration'

INPUT: 'You just talked about having this record day. I think you said last Saturday and I think you're also maybe appropriately cautioning about third-quarter trends and variance and Delta. So just kind of put those things together. Record day still sounds like things are good, but I'm sure there are some warnings out there. Can you be more specific about you're seeing that's causing you to be just a little more cautious about Q3 and the back half of the year? Thanks.'

OUTPUT: 'So just kind of put those things together.': 'elaboration', 'Can you be more specific about you're seeing that's causing you to be just a little more cautious about Q3 and the back half of the year?': 'clarification'

INPUT: 'With flexible dates rolling out, how has the booking window changed? Are there more instances of last minute stays coming into play?'

OUTPUT: 'how has the booking window changed?': 'explanation', 'Are there more instances of last minute stays coming into play?': 'data'

INPUT: 'Murray, can I ask on the gas contract. Does that mean that customers are reluctant to term out these sort of prices?'

OUTPUT: 'Does that mean that customers are reluctant to term out these sort of prices?': 'confirmation'

INPUT: 'I have a couple of questions on investments and spending. I guess first off, what gives you the confidence that you can moderate sales and marketing spend in the second half without impacting bookings or listing levels? Is that still the plan?'

OUTPUT: 'what gives you the confidence that you can moderate sales and marketing spend in the second half without impacting bookings or listing levels?': 'justification', 'Is that still the plan?': 'confirmation'

INPUT: 'I have a couple on the long-term stays of 28 days or more. You guys mentioned that it was roughly 20% of nights spent in 3Q. But I think in the first quarter, it was around 24%. Was this mostly seasonality that it declined slightly since then?'

OUTPUT: 'Was this mostly seasonality that it declined slightly since then?': 'explanation'

INPUT: 'And then the second question, if I could. Just in the letter you talked about in relation to APAC,

just the short-term rental restrictions as being somewhat of a drag on the recovery in that region. Are those new restrictions versus 2019? And how significant are they? Do you expect them to be a permanent drag?

OUTPUT: 'Are those new restrictions versus 2019?': 'clarification', 'And how significant are they?': 'data', 'Do you expect them to be a permanent drag?': 'opinion'

Following, the text you need to analyze. INPUT: *text*.

Return ****only**** the JSON output.

B Linguistic features annotation errors

Figure 2 reports the confusion matrices referring to the best-performing models (F_1 assessment metric) in the automatic annotation task for features discourse regulators, prefaces, question types, and request types. As proposed by Ortman (2022), the horizontal axis presents the predicted labels and the vertical axis the gold-standard (target) labels. Each pairwise intersection of the matrix represents an annotation error. In case the column and row labels are different, the number corresponds to the sum of pure labeling and labeling + span-length errors; if the column and row labels are the same, the reported number are pure span-length errors. The last column and row indicate false negatives and false positives respectively.

Target label	Predicted label									
	recipient	theme	enumeration	counting	greeting	acknowledgement	agreement	inside_comment	outside_comment	Ø (FN)
recipient	2	0	0	0	0	0	0	0	0	6
theme	0	4	1	0	0	0	0	0	0	15
enumeration	0	0	0	0	0	0	0	0	0	8
counting	0	1	1	2	0	0	0	0	0	14
greeting	0	0	0	0	2	0	0	0	0	16
acknowledgement	0	0	0	0	0	5	0	0	0	14
agreement	0	0	0	0	0	0	0	2	0	2
inside_comment	0	0	0	0	0	14	4	0	0	14
outside_comment	0	0	0	0	0	0	0	1	0	6
Ø (FP)	16	26	6	11	12	7	5	13	4	0

(a) Discourse regulators: gemini-2.0-flash

Target label	Predicted label				
	fact	reported speech	opinion	meta-discourse	Ø (FN)
fact	108	17	26	3	105
reported speech	5	94	3	1	35
opinion	18	7	60	2	54
meta-discourse	4	2	3	7	23
Ø (FP)	239	192	132	36	0

(b) Prefaces: gpt-4o

Target label	Predicted label			
	open	yes/no	closed-list	Ø (FN)
open	23	14	0	18
yes/no	0	11	1	7
closed-list	0	1	5	1
Ø (FP)	1	3	1	0

(c) Question types: gemini-1.5-pro

Target label	Predicted label							
	clarification	commitment	confirmation	data	elaboration	explanation	justification	opinion
clarification	6	1	4	1	2	0	0	0
commitment	0	0	0	0	0	0	0	0
confirmation	2	0	6	0	0	0	0	3
data	3	0	4	5	0	0	0	4
elaboration	1	0	0	5	6	4	0	1
explanation	1	0	0	0	1	2	0	1
justification	1	0	0	0	0	0	0	0
opinion	3	0	3	5	3	3	0	8
Ø (FP)	7	1	13	2	5	1	0	0

(d) Request types: claude-3.7-sonnet

Figure 2: Confusion matrices reporting results for the best-performing models in annotating discourse regulators, prefaces, question types, and request types.

C Automatic annotation of linguistic features

	discourse regulators		prefaces		question types		request types	
	F_1	Γ	F_1	Γ	F_1	Γ	F_1	Γ
gpt-4o	<u>0.52</u>	0.12	0.66	0.29	<u>0.41</u>	<u>0.35</u>	0.24	<u>0.32</u>
gemini-1.5-pro	0.45	-0.16	0.62	0.25	0.52	-0.64	0.26	0.33
gemini-2.0-flash	0.55	<u>0.11</u>	<u>0.65</u>	0.22	0.35	0.38	<u>0.29</u>	<u>0.32</u>
claude-3.7-sonnet	0.50	-0.14	0.61	<u>0.27</u>	0.40	-0.07	0.31	0.04
human baseline (IAA)	—	<u>0.78</u>	—	<u>0.76</u>	—	<u>0.82</u>	—	<u>0.72</u>

Table 4: Feature annotation performance across models (F_1 , Γ), compared against human IAA (Γ). Among LLMs: for each feature annotation task, for each metric, best results are in bold; second best underlined. Human annotation overperforms any model.

D Correlations between question origin and features

	Spearman ρ	p -value
Prefaces		
number	-0.676	<.001
length	-0.671	<.001
type		
<i>fact</i>	-0.435	<.001
<i>opinion</i>	-0.352	<.001
<i>rep. speech</i>	-0.346	<.001
<i>meta-discourse</i>	-0.204	0.042
Question types		
<i>open</i>	-0.094	0.354
<i>polar</i>	-0.169	0.093
<i>closed-list</i>	-0.295	0.003
Request types		
<i>clarification</i>	-0.204	0.042
<i>confirmation</i>	-0.333	0.001
<i>data</i>	+0.020	0.842
<i>elaboration</i>	-0.165	0.101
<i>explanation</i>	-0.252	0.011
<i>justification</i>	-0.143	0.156
<i>opinion</i>	-0.238	0.017
Discourse regulators		
<i>recipient</i>	-0.500	<.001
<i>theme</i>	-0.456	<.001
<i>greeting</i>	-0.547	<.001
<i>agreement</i>	-0.143	0.156
<i>enumeration</i>	-0.295	0.003
<i>counting</i>	-0.229	0.022
<i>acknowledgment</i>	-0.655	<.001
<i>inside comment</i>	-0.101	0.320
<i>outside comment</i>	-0.138	0.172
Vocabulary entropy	-0.759	<.001
Flesch-Kincaid	+0.597	<.001
Dale-Chall	+0.634	<.001
Word count	-0.769	<.001
Sentence count	-0.855	<.001
Type-token ratio	+0.739	<.001
Filler word count	-0.722	<.001
Interjection count	-0.498	<.001
Stopword count	-0.781	0.051
NER (person)	-0.312	0.002
Question count	-0.518	<.001
Assertion count	-0.733	<.001
Mean question length	+0.316	0.001
Mean assertion length	-0.654	<.001

Table 5: Spearman correlations between question origin (negative value: correlates with natural origin; positive value: correlates with synthetic origin) and features selected in phase one of the study. In bold the statistically significant values of effect size.

E Prompt for synthetic questions generation

You are an expert in financial communication.

Listed companies hold quarterly meetings with financial analysts called earnings conference calls (ECCs). ECCs encompass a first part of “prepared remarks” and a second part of “Q&A”. The questions asked during the Q&A session are primarily on the topics touched during the prepared remarks, which in turn recollect the main points of the quarterly financial disclosure that was published not much earlier than the ECC.

Here an example of complete ECC (prepared remarks followed by the Q&A session). Please revise it carefully.

Example: *example*

Now, I’ll give you some prepared remarks. Information about the company the transcript refers to, alongside the quarter and the year, are included in the text. You must act as a financial analyst and produce 20 different questions that could have been asked in the subsequent Q&A session by attendees. Refer to the contents of the prepared remarks but also, as a financial analyst would do, you can integrate with external knowledge that you may have about the company and its current situation. Of course, the cutoff date is the day of the event: since you’re pretending that you’re attending it, you cannot have any knowledge about the future.

Your response must be in JSON format, UTF-8 encoded, following the template “Question 1”: “text of question 1”, “Question 2”: “text of question 2” ...

This is a transcript of prepared remarks from an ECC. Analyze it carefully and produce 20 novel question as professionals (financial analysts) would ask in this occasion. Output the correctly formatted JSON file **only**.

Prepared remarks: *prepared remarks*

NOTES ON THE PROMPT:

The safeguarding instruction about ignoring pretraining information following the ECC cutoff date is probably not sufficient to ensure what was requested. However, it was only intended to make the model aware that temporal mismatches should not be obvious at the surface level or give rise to logical contradictions, which our respondents would have noticed. Respondents presumably do not have sufficient financial literacy to detect factual inconsistencies, so we are not concerned about them.

The risk of the model “generating” a real question verbatim, thus potentially duplicating a natural question already present in the dataset, is not a possibility we envisioned: LLMs invariably rephrase, even when instructed not to, and that already qualifies as a synthetic turn for the purposes of our study.

F Further information about questionnaires building

Natural questions were randomly sampled from 22 ECCs (2 for each of the stock market sectors, according to The Global Industry Classification Standard (GICS)¹⁰) from years 2020-2024. Synthetic questions were generated by gpt-4o through one-shot prompting of an entire ECC transcript, randomly selected at each API call among a pool of 11 (each representing a stock market sector, from the same timeframe as the natural questions), and including the authentic prepared remarks about which the synthetic questions need to be generated. Prepared remarks are the first part of an ECC, in which quarterly results of the company are presented and commented by corporate representatives. As an approximation, we can assume that also in a natural setting, questions asked in the Q&A session are *prompted* by such preceding part. (Crawford Camiciottoli, 2010)

G Questionnaire setup

The participants on Prolific were instructed as follows:

“Given a question that was asked in the context of a financial earnings call, your task is to discern whether the tone with which the question was formulated is – according to your own personal understanding – ‘professional’, ‘somewhat professional’ or ‘not professional’.

You are not expected to have any prior knowledge in finance.

You will be requested to rate 10 questions in total; you are allowed to opt out from this study at any moment if you don’t feel like completing the task, and your partial response will not be recorded.

Your responses will be stored and used anonymously for research purposes.

Your responses will be employed ethically without any intent to discriminate on the basis of the “professionalism” rating that an item may receive.”

Then, for each item:

“Question: (*question text*).

How would you rate this question? Is it ‘professional’, ‘somewhat professional’ or ‘not professional’?”

¹⁰<https://www.msci.com/our-solutions/indexes/gics>

H Regression models

	coef	std err	<i>p-value</i>
Vocabulary entropy	+0.2301	0.027	<.001
Flesch-Kincaid	+0.0001	0.004	0.973
Dale-Chall	+0.0442	0.011	<.001
Word count	-0.0017	0.002	0.438
Sentence count	-0.0028	0.008	0.727
Type-token ratio	+0.9810	0.170	<.001
Filler word count	-0.0130	0.005	0.013
Interjection count	-0.0411	0.023	0.080
Stopword count	+0.0040	0.003	0.203
NER (person)	-0.0285	0.017	0.090
Question count	+0.0130	0.009	0.161
Assertion count	-0.0158	0.007	0.030
Mean question length	+8.41e-05	<.001	0.754
Mean assertion length	-0.0003	<.001	0.357

Table 6: Best regression model (three way classification, NLP features); $R^2 = 0.983$. In bold the statistically significant values. Positive values align with positive perceived professionalism.

	coef	std err	p-value
Prefaces			
length	-0.0002	<.001	0.652
type			
<i>fact</i>	-0.0569	0.037	0.649
<i>opinion</i>	+0.0423	0.035	0.225
<i>rep. speech</i>	+0.0038	0.034	0.911
<i>meta-discourse</i>	-0.0039	0.043	0.929
Question types			
<i>open</i>	+0.0458	0.032	0.150
<i>polar</i>	-0.0233	0.031	0.458
<i>closed-list</i>	+0.0188	0.034	0.585
Request types			
<i>clarification</i>	+0.0141	0.038	0.712
<i>commitment</i>	-0.0235	0.049	0.636
<i>confirmation</i>	-0.0383	0.034	0.265
<i>data</i>	-0.0212	0.030	0.479
<i>elaboration</i>	+0.0283	0.032	0.381
<i>explanation</i>	-0.0130	0.036	0.721
<i>justification</i>	-0.0571	0.082	0.487
<i>opinion</i>	-0.0259	0.030	0.388
Discourse regulators			
<i>recipient</i>	+0.0012	0.034	0.971
<i>theme</i>	-0.0151	0.033	0.649
<i>greeting</i>	-0.0227	0.036	0.534
<i>agreement</i>	-0.0701	0.055	0.202
<i>enumeration</i>	-0.0420	0.042	0.322
<i>counting</i>	-0.0176	0.040	0.656
<i>acknowledgment</i>	-0.0439	0.043	0.313
<i>inside comment</i>	-0.0378	0.042	0.368
<i>outside comment</i>	+0.0388	0.041	0.341
Vocabulary entropy	+0.2519	0.030	<.001
Flesch-Kincaid	+0.0009	0.005	0.834
Dale-Chall	+0.0400	0.011	0.001
Word count	-0.0012	0.002	0.595
Sentence count	+0.0032	0.009	0.722
Type-token ratio	+0.9154	0.177	<.001
Filler word count	-0.0115	0.005	0.036
Interjection count	-0.0273	0.025	0.275
Stopword count	+0.0026	0.003	0.423
NER (person)	-0.0396	0.019	0.035
Question count	+0.0140	0.011	0.191
Assertion count	-0.0108	0.008	0.187
Mean assertion length	-0.0005	<.001	0.160

Table 7: Second best regression model (three way classification, combined features optimized); $R^2 = 0.983$. In bold the statistically significant values. Positive values align with positive perceived professionalism.