



Bridging form and function: An exploratory LLM approach toward automated retrieval of directive speech acts in spoken English

Guying Zhou^a, Jiajin Xu^{b,*}

^a School of German Studies, Beijing Foreign Studies University, Beijing, China

^b National Research Center for Foreign Language Education, Beijing Foreign Studies University, Beijing, China



ARTICLE INFO

Article history:

Received 12 January 2026

Received in revised form 3 June 2026

Accepted 16 July 2026

Keywords:

Corpus pragmatics

Automated retrieval

Directive speech acts

Form-function mismatch

LLM-based

ABSTRACT

Automated retrieval of speech acts in naturally occurring conversations has yet to be systematically explored within linguistics, particularly for directive speech acts—characterized by a pronounced mismatch between linguistic forms and their communicative functions. Based on dassBNC2014, a self-compiled dataset, this study explored a quantitative methodology toward the automated retrieval of speech acts using large language models (LLMs). We compared three retrieval methods that exploit linguistic and contextual information to different degrees: (1) lexical search using conventionalized linguistic markers; (2) feature-based support vector machine (SVM) classifiers drawing on lexical, syntactic, and surface-level contextual features; and (3) GPT-4o (a general-purpose model and a task-specific fine-tuned variant) used as directive classifiers. Retrieval effectiveness is evaluated using precision, recall, and F2-score, and is complemented by an error analysis of false positives and false negatives. The results indicate that lexical search yields very limited performance, and that incorporating richer linguistic and surface-level contextual features improves recall, but only modestly. By contrast, the fine-tuned LLM achieved the best overall performance. This study demonstrates the methodological potential of LLMs for bridging form and function in corpus-based pragmatic research and provides an empirical foundation for scaling speech act retrieval to other categories.

© 2026 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

1. Introduction

Human language extends beyond being a mere code for exchanging information. One of the most compelling aspects of speech act theory (Austin, 1962; Searle, 1969) lies in the fact that, despite the inexhaustible diversity of linguistic forms, utterances in human communication can be broadly classified into basic categories contingent upon their illocutionary force. Beyond linguistics, this very concept, often referred to as “dialog act” (see Stolcke et al., 2000; Ritter et al., 2010; Bunt, 2022; Jurafsky and Martin, 2024), has also become foundational in natural language processing (NLP), where it plays an

* Corresponding author.

E-mail address: xujiajin@bfsu.edu.cn (J. Xu).

important role in dialog systems, conversational modeling, and discourse-level interpretation. In both linguistics and NLP, the reliable retrieval of speech acts is a necessary foundation for further analysis and system development.

However, retrieving speech acts from naturally occurring conversations remains a considerable challenge (see Jucker et al., 2008: 273; Rühlemann, 2010: 289). Speech acts are not primarily defined in terms of linguistic form, but rather in terms of their communicative functions. There is often a substantial gap between functions and forms. As Levinson (2017: 203) notes, speech acts are “not for the most part simply or directly coded in the linguistic form.” This lack of direct formal encoding, frequently discussed as the form-function mismatch (see Kohnen, 2015: 54; Milà-Garcia, 2018: 265; Rühlemann, 2019: 19), complicates efforts to identify speech act realizations.

In corpus linguistics, two major methodological approaches have emerged: 1) The form-to-function approach, which begins with predetermined linguistic forms and assumes that these forms map onto a particular speech act function. Three main strategies are commonly used, namely illocutionary force indicating devices (IFIDs), syntactic patterns, and meta-communicative expressions (Jucker and Taavitsainen, 2014; Jucker, 2018). IFID-based retrieval targets lexical items conventionally associated with particular speech acts, for example *sorry*, *excuse* for apologies, *thank* for gratitude, *please* for requests, and *hello* or *bye* for greetings and farewells (Aijmer, 1996; Deutschmann, 2003, 2006; Schauer and Adolphs, 2006; Wong, 2010; Jucker, 2017; Su 2017; Arizavi and Choubasaz, 2019; Degenhardt and Bernaisch, 2022). Other studies have transformed recurrent syntactic patterns into search strings, as in Jucker et al.'s (2008) retrieval of compliments, or used metacommunicative labels such as *compliment* to identify relevant instances in COHA and COCA (Jucker and Taavitsainen, 2014). Supplementary techniques include collocational filtering (Lutzky and Kehoe, 2017) and semiautomatic retrieval with manual checks (Su, 2017). This approach works best for highly conventionalized speech acts such as apologies, which can reach recall rates of nearly 98% (Kitao, 2012). In contrast, less-conventionalized acts, such as requests, achieve far lower recall (76.8%; Kitao and Kitao, 2012). Indirect speech acts, particularly unconventional ones, remain difficult to capture because they lack stable surface cues. Consequently, studies using this approach tend to focus on conventionalized forms and cannot easily reveal new form-function correspondences. 2) The function-to-form approach, however, reverses the process by identifying speech acts through manual context-sensitive inspection. This “conversation-analytical” method (Jucker, 2009) is labor-intensive and therefore typically applied to small corpora. Nevertheless, it has produced influential work across genres, including academic discourse (Garcia, 2004), business meetings (Bilbow, 2002), workplace conversations (Koester, 2002), service encounters (Gagne, 2018), diachronic directive studies (Kohnen 2008), and diverse conversational and CMC datasets (e.g., Merrison et al., 2012; Bell, 2020; Burnette and Calude, 2022). Its chief advantage lies in its ability to capture the full pragmatic spectrum of a speech act, including indirect and weakly conventionalized forms.

Advances in NLP have demonstrated that the automated retrieval of speech acts is technically feasible. Early dialog act classification relied heavily on feature engineering. Two modeling strategies were common: multiclass classification, such as logistic regression, Support Vector Machines (SVMs), Naïve Bayes; and sequence labelling, such as Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs). The latter capturing dependencies across turns. Representative studies include Bayesian networks (Keizer 2001), HMM-based systems (Stolcke et al., 2000), N-gram models (Louwerse and Crossley, 2006), multimodal SVM/HMM approaches (Surendran and Levow, 2006), and SVM-based studies on social media data (Zhang et al., 2011; Ribeiro et al., 2015; Qadir and Riloff, 2011). Although these methods achieved moderate accuracy, their scalability is limited by domain-specific datasets and their dependence on handcrafted features. Deep learning (DL) substantially reduces the dependence on manual features and improves the performance on benchmark corpora such as SwDA and MRDA. Studies combining CNNs with context modeling (Lee and Dernoncourt, 2016), hierarchical Bi-LSTMs (Kumar et al., 2018), attention mechanisms (Ortega and Vu, 2017), and context-aware self-attention (Raheja and Tetreault, 2019) have consistently outperformed traditional ML models. Most recently, pretrained language models (ELMo, BERT, RoBERTa, GPT-2, and DialoGPT) have become the state-of-the-art. Evaluations by Duran et al. (2023) demonstrated that transformer-based models outperform both traditional ML and DL approaches across major datasets, offering a streamlined pipeline and superior accuracy, albeit at a higher computational cost.

Taken together, research in both corpus linguistics and NLP has sought to narrow the gap between linguistic form and communicative function, and to develop more reliable procedures for retrieving speech acts. Despite the growing interest in computational methods, corpus-based pragmatic research continues to rely predominantly on manual identification or surface-form searches. Methodological development in speech act retrieval has lagged behind, and the potential for systematic integration between corpus linguistics and NLP remains underexploited. As a result, evaluations on automated retrieval of speech acts have generally remained cautious in the field. For example, Nagy et al. (2025) case study notes that indirect and implicit speech acts cannot be identified automatically and that their analysis continues to depend on pragmatic intuition and qualitative contextual interpretation. Jucker et al. (2008: 273) argued that corpus-based methods may be ill-suited for the study of speech acts due to a lack of direct and stable formal encoding. Against this backdrop, the present study explores a quantitative methodology toward automated retrieval of speech acts, with particular emphasis on the potential of LLM-based methods. Thus, the following research question is addressed:

Among retrieval methods that rely on lexical, syntactic, and contextual features to varying degrees, which approach achieves the best performance in retrieving directive speech acts?

We hypothesized that the retrieval performance would improve systematically as the methods incorporate richer linguistic and contextual information. Approaches relying solely on surface-level cues are expected to exhibit limited precision and recall. The integration of syntactic and other structured linguistic features, and surface-level contextual information

should yield measurable gains, and LLM-based retrieval is hypothesized to perform most effectively by exploiting broader contextual dependencies that enable a more robust mapping between linguistic form and communicative function.

To address this research question and test this hypothesis, we conducted three retrieval experiments using a self-compiled dataset of directive speech acts. Directive speech acts provide a particularly stringent test case because of their pronounced form-function mismatch and frequent realization through indirect or structurally non-conventional forms. Therefore, the demonstration of robust retrieval performance in this domain suggests the potential applicability of the proposed approach to other speech act types.

2. Data and methodology

2.1. Corpus

The corpus used in this study was the Directive Annotated Subset of Spoken BNC2014 (dassBNC2014),¹ created specifically for the present research. As the name suggests, it was derived from the Spoken BNC2014 (Love et al., 2017). Sixty-one transcripts of recorded conversations were selected through stratified sampling across five speaker-relationship categories. Together, they represent 40 h, 14 min, and 44 s of conversational speech, corresponding to 470,024 tokens and 72,106 sub-utterances.² Of these, 2333 were manually annotated as directives.³

We further divided the dataset into training, test, and validation subsets using a stratified approach. As shown in Table 1, the training, test, and validation sets comprised 61.29%, 18.34%, and 20.36% of the annotated directives, respectively.⁴

Table 1
Overview of training, test and validation set of dassBNC2014

		Files	Length	Sub-utterances	Directives
Training set (61.29%)	Close family	14	09:43:17	17,116	602
	Friends	7	07:03:47	12,485	455
	Colleges	3	02:07:57	3633	172
	Acquaintances	4	03:10:24	6169	177
	Strangers	4	02:25:46	3830	24
	SUM	32	24:31:11	43,233	1430
Test set (18.34%)	Close family	5	02:25:33	3634	142
	Friends	4	01:45:54	2592	90
	Colleges	5	01:46:24	3515	118
	Acquaintances	2	00:53:53	1773	38
	Strangers	2	00:33:42	1879	40
	SUM	18	07:59:08	13,393	428
Validation set (20.36%)	Close family	2	02:20:18	5447	213
	Friends	3	02:20:59	3719	154
	Colleges	2	01:25:36	3591	71
	Acquaintances	3	01:18:40	2060	24
	Strangers	1	00:18:52	663	13
	SUM	11	07:44:35	15,480	475
TOTAL		61	40:14:44	72,106	2333

Based on this empirical foundation, we conducted retrieval experiments and evaluated model performance against the manual annotations, which served as the “gold standard” for directive identification.

¹ Our dataset, dassBNC2014, is available on GitHub (<https://github.com/buyingo/dassBNC2014>).

² In the original Spoken BNC2014, the speaker turn is treated as the basic unit of the conversational data. For speech act annotation, however, this unit is too coarse-grained. In the present dataset, we therefore adopt the sub-utterance as the basic annotation unit, defined as “syntactically and pragmatically independent segments within speaker turns.” Based on annotation trials and comparison with two alternative unit types, we found sub-utterances to be the most stable and the most suitable for speech act annotation. The segmentation of sub-utterances was carried out using GPT-4o-mini with few-shot prompting. Further details can be found in Zhou (2025) and Zhou and Xu (in press).

³ Directives are defined as utterances in which the speaker intends the hearer to perform an action, regardless of whether the intention is sincere or whether the action is actually carried out. We follow the international standard for speech act annotation (ISO 24617-2, 2020), which distinguishes between two main general-purpose communicative functions: information transfer (e.g., providing or requesting information) and action discussion (e.g., initiating or negotiating actions). Directives are treated as instances of the latter, since they are oriented toward eliciting concrete actions from hearers.

⁴ Given the significant class imbalance in our directive-annotated dataset, where positive cases (i.e., utterances annotated as directive speech acts) accounted for only about 3% of all sub-utterances, we performed undersampling to construct a balanced training set. Specifically, the training data included all positive examples (1430 instances), 333 false positives identified during the annotation process, and 1097 randomly selected negative instances, all from the training set. This yielded a curated dataset of 2860 sub-utterances, evenly distributed between positive and negative cases.

2.2. Retrieval methods

This section introduces the three retrieval methods compared in the present study. Lexical search relies solely on surface cues such as lexical items and basic syntactic patterns. SVM-based feature engineering incorporates a broader range of linguistic features together with superficial contextual features. LLM-based retrieval integrates a deeper level of contextual information using the GPT-4o.

2.2.1. Lexical search

A lexical search provides a form-based baseline for automatic retrieval of directives. In line with previous studies, we implemented three types of form-based cues (see [Jucker and Taavitsainen, 2014](#), [Jucker, 2018](#)):

- 1) Illocutionary force indicating devices (IFIDs), such as *please*, performative verbs (*ask, request, order, recommend*), and imperative verb forms;
- 2) Syntactic patterns reflecting directive intent, including second-person and modal constructions (*you have to/you should/would you mind/can you*), first-person request frames (*I want you to/I'd like you to*), and inclusive forms (*we need to/can we*).
- 3) Metacommunicative expressions that explicitly label directive actions (*request, requirement, demand, suggestion, advice, etc.*).

The retrieval is implemented in Python, using spaCy ([Honnibal and Montani, 2017](#)) for POS tagging and dependency parsing. Because the goal was to capture as many directive speech acts as possible, we prioritized recall over precision by employing relatively flexible search rules.

2.2.2. Support vector machine (SVM)

Support vector machines (SVMs) have long been recognized as robust and widely employed machine learning approaches ([Cortes and Vapnik, 1995](#); [Vapnik, 1998](#)). More recently, the application of SVM has begun to appear in linguistic research on speech-act retrieval (e.g. [John et al., 2019](#)).

To train the SVM classifiers, we constructed a feature set that captured the lexical, syntactic, semantic, and superficial contextual properties of the target sub-utterance (see [Table 2](#)). The first classifier draws solely on linguistic information. It uses 661 lexical features, 318 syntactic features, and 384 semantic dimensions, resulting in 1363 features in total. This model represents a maximally enriched form-based approach without contextual input. The second classifier extended this feature set by adding 807 surface-level contextual features, yielding 2170 features overall. Both the SVM experiments were implemented using scikit-learn ([Pedregosa et al., 2011](#)).

Table 2
Feature set used in the SVM classifiers.

Feature group	Feature type	Description
(i) Lexical	N-grams	Uni-/bi-/trigrams of the target sub-utterance selected via chi-square filtering
	Personal pronouns	First-, second-, third-person pronoun indicators
	Modal verbs	<i>can, may, must, shall, will, might, should, would, and could</i>
	Wh-words	<i>who, when, where, what, which, and how</i>
	Please	Politeness marker <i>please</i>
	Speech act verbs	Verb list drawn from Searle (1979) , Bach and Harnish (1979) , and Wierzbicka (1987)
(ii) Syntactic	Metacommunicative nouns	Nominalized forms of speech act verbs
	Number of verbs	POS-based counts of verbs
	Number of adjectives	POS-based counts of adjectives
	Dependency sub-trees	Frequent sub-trees (length 1–2) drawn via spaCy, chi-square selected
	Tense	Binary encoding of present, past, and future tense
	Directive patterns	Regex patterns
	You + infinitive	Pattern detection via spaCy
	Sentence-initial forms	Starting with a modal or a verb
	Wh-about patterns	<i>what about, how about</i> constructions
	Question mark	Binary indicator of interrogative form
(iii) Semantic	Sentence embeddings	384-Dim vectors representing sentence meaning
(iv) Contextual	Context n-grams (± 5)	N-grams from five preceding and following sub-utterances
	Neighboring question	Binary indicator of whether the next sub-utterance contains a question mark
	Speaker gender	Male/female/NA
	Speaker relationship	Five relationship categories encoded as binaries

2.2.3. LLM-based retrieval

Pre-trained language models have achieved state-of-the-art performance in tasks such as speech act retrieval, outperforming traditional machine learning and early deep learning methods (see Duran et al., 2023). The latest developments include large-scale generative language models commonly referred to as large language models (LLMs). In this study, we assessed the performance of a large language model in the task of retrieving directive forces, which represents a further step toward the function-oriented end of the continuum. By incorporating deeper contextual information, the LLM provides a holistic approach to speech act identification and transcends surface-level linguistic features to capture functional and contextual subtleties.

Specifically, we used OpenAI's GPT-4o in two configurations as demonstrated in Fig. 1: the base model, which served as a general-purpose benchmark, and the fine-tuned version, GPT-4o-Directives, which was optimized for directive identification.

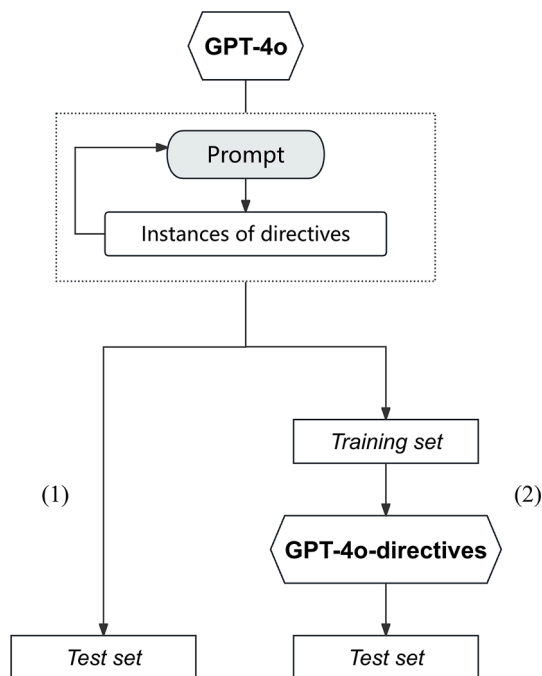


Fig. 1. Prompt-based learning and full-model fine-tuning.

In prompt-based fine-tuning, we iteratively tested potential prompts on sample dialog fragments.⁵ An effective prompt was then finalized and applied to the batch annotation using the OpenAI API. The model performance was assessed using the test set. In supervised fine-tuning, annotated training data are used to create a task-specific variant, GPT-4o-Directives, which is evaluated on the test set. Both approaches incorporate context by providing the target sub-utterance together with the preceding and following 10 sub-utterances.⁶

2.3. Workflow

In this study, we aimed to explore which retrieval method most effectively identifies directive speech acts. We conducted three retrieval experiments. The workflow is illustrated in Fig. 2 below.

⁵ Before conducting batch retrieval via the OpenAI API, we experimented with a range of prompts in the GPT-4o chat box, including both zero-shot and few-shot variants. This iterative process led to the prompt ultimately adopted in the present study, which is reproduced in Appendix.

⁶ A context window of ten sub-utterances on either side of the target sub-utterance was chosen as a practical compromise. In most cases, it provides enough contextual information, typically including the relevant response from the conversational partner, while keeping both cost and potential losses in model performance at a manageable level.

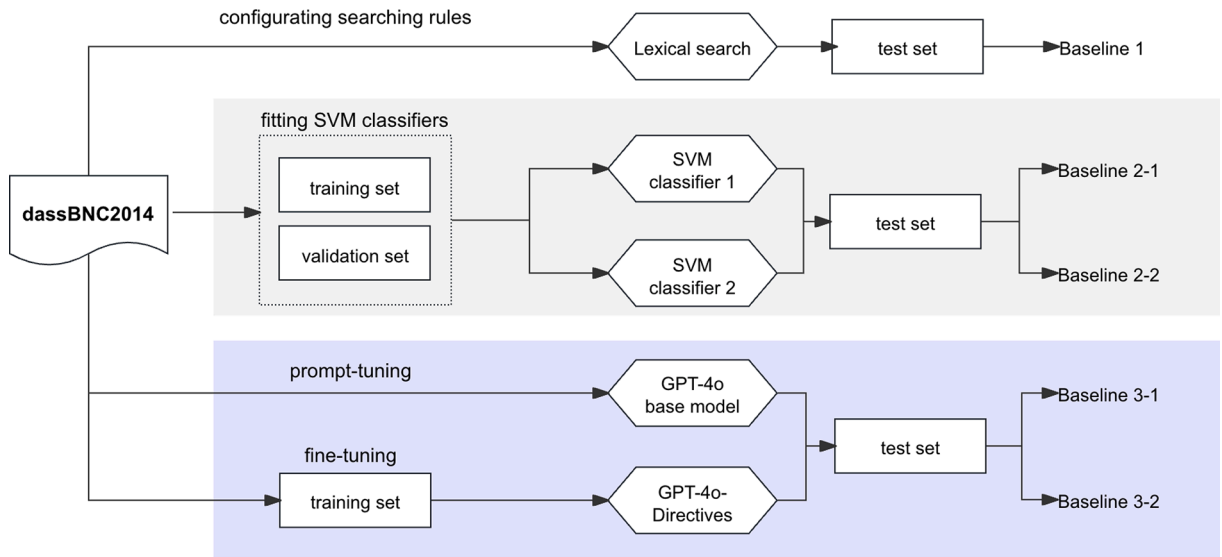


Fig. 2. Research workflow.

The first approach, lexical-search-based retrieval, yields Baseline 1. It assumes that directive force is, at least to some degree, encoded in linguistic form, yielding Baseline 1 as a strictly form-based benchmark.

The second approach (shaded in light gray in the figure above), SVM-based retrieval extends the lexical search by incorporating a wider range of linguistic features along with limited, superficial contextual cues. We trained two SVM classifiers and evaluated each on the test set, producing Baseline 2-1 and Baseline 2-2. These baselines assess the extent to which an expanded set of linguistic features can capture directive illocutionary forces, and test whether even minimal contextual information improves retrieval performance.

The third approach, the LLM-based retrieval (shaded light blue), incorporates substantially deeper contextual information using the GPT-4o model. We evaluated both the base model (Baseline 3-1), which served as a general-purpose benchmark, and a fine-tuned variant, GPT-4o-Directives (Baseline 3-2), to compare general-purpose and task-adapted performances.

Altogether, these experiments yielded three sets of baselines, each evaluated in terms of *precision*,⁷ *recall*,⁸ and *F2-score*,⁹ enabling a systematic comparison of methods that rely on lexical, syntactic, and contextual information to different degrees.

3. Performance metrics of different retrieval methods

In this section, we present the results of three retrieval experiments. For each retrieval, we report the performance metrics, calculated on the test set described in Section 2.1. The test set contains 13,393 sub-utterances, of which 334 empty entries were removed. Among the remaining 13,059 sub-utterances, 428 were manually annotated as directive speech acts, resulting in a highly imbalanced distribution in which positive instances accounted for only 3% of the data. Positive cases were labeled as 1 (directives) and negative cases as 0 (non-directives).

3.1. Baseline 1

A sub-utterance is classified as a directive if it satisfies any of the predefined rules, as illustrated in Section 2.2. The performance of the lexical search on the test set is shown in Table 3.

⁷ Precision indicates the proportion of retrieved instances that are correct.

⁸ Recall indicates the proportion of all relevant instances that are successfully retrieved.

⁹ The F2-score is a recall-oriented variant of the F-score that gives greater weight to recall than to precision. It is useful in tasks where missing relevant positive cases is considered more problematic than retrieving some false positives. Formally, $F = \left(1 + 2^2\right) \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$.

Table 3

Confusion matrix for lexical-search-based retrieval on test set.

		Lexical search	
		0	1
AN00	0	11,435 (TN)	1196 (FP)
	1	198 (FN)	230 (TP)

In the test set, 11,435 sub-utterances were correctly identified as non-directives, representing true negatives (TN). Conversely, 1196 sub-utterances were incorrectly classified as directives, constituting false positives (FP). Among the manually identified directive speech acts, 198 instances were missed by the retrieval method, accounting for false negatives (FN). However, this method successfully retrieved 230 directives representing true positives (TP).

Based on the confusion matrix, we calculated Baseline 1. A precision of 16.13% shows that many non-directives are incorrectly flagged as directives, whereas a recall of 53.74% indicates that a substantial number of true directives remain unretrieved. The F2-score of 0.3665 further demonstrated the limited effectiveness of the method. Overall, these results show that lexical and syntactic matching alone is insufficient for reliably identifying directive speech acts. Baseline 1 should therefore be understood primarily as a transparent and interpretable reference point against which the performance of Baselines 2 and 3 can be evaluated.

3.2. Baseline 2-1 and 2-2

Following hyperparameter tuning on the validation set, the optimal parameters for our dataset were determined to be {'C': 1, 'kernel': 'rbf'}¹⁰. Using these parameters, we trained SVM classifiers 1 and 2 on the balanced training dataset. The performance of both classifiers was then evaluated using the test set.

As shown in Table 4, SVM classifier 1 correctly identified 11,281 sub-utterances as TNs. However, 1350 sub-utterances were incorrectly classified as FPs. Among the manually annotated directive speech acts, 119 instances were not detected by the classifier, accounting for FNs. Meanwhile, the classifier successfully retrieved 309 as TPs.

Table 4

Confusion matrix for SVM classifier 1 on test set.

		SVM classifier 1	
		0	1
AN00	0	11,281 (TN)	1350 (FP)
	1	119 (FN)	309 (TP)

The precision was 18.63%, indicating frequent misclassification of non-directives as directives and thus many false positives. The recall was 72.20%, showing that while the classifier retrieves some true directives, it missed a substantial number. The F2-score is 0.4583.

Table 5 shows that SVM classifier 2 successfully identified 11,067 sub-utterances as TNs. However, 1564 sub-utterances were incorrectly classified as FPs. Among the manually annotated directive speech acts, 107 instances were missed, accounting for FNs. Nevertheless, the classifier correctly identified 321 directives as TPs.

Table 5

Confusion matrix for SVM classifier 2 on test set.

		SVM classifier 2	
		0	1
AN00	0	11,067 (TN)	1564 (FP)
	1	107 (FN)	321 (TP)

Classifier 2 achieved a precision of 17.03%, which showed that incorporating superficial contextual information did not reduce the misclassification of non-directives as directives, resulting in numerous false positives. A recall of 75.00% indicates

¹⁰ The parameter C controls the trade-off between maximizing the margin and minimizing classification errors, while the RBF kernel allows the model to capture non-linear decision boundaries.

a modest improvement in the retrieval of true directives, although some instances are still missed. The F2-score of 0.4462 further demonstrates the limited effectiveness.

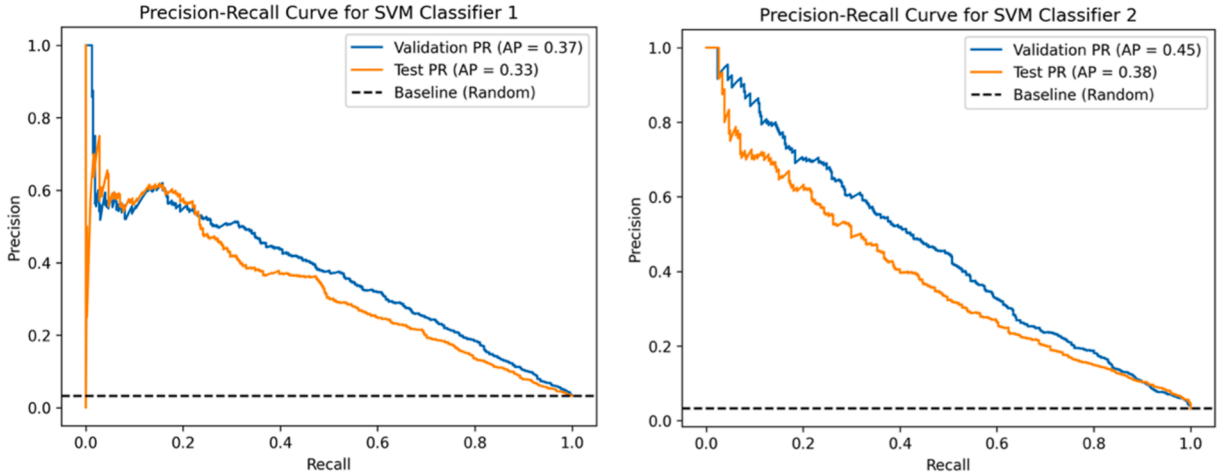


Fig. 3. Precision-recall curve for SVM classifier 1 and 2.

Fig. 3 presents the precision–recall curves for SVM classifiers. The average precision (AP) of classifier 1 is 0.37 on the validation set and 0.33 on the test set. For classifier 2, the average precision (AP) is 0.45 on the validation set and 0.38 on the test set. Both are well above the random baseline of 0.03 imposed by class imbalance. Although the test set shows a slight drop in performance compared to the validation set, the consistent shapes of the validation and test curves suggest that the model does not suffer from significant overfitting.

Overall, SVM classifier 1 represents a comprehensive exploration of the linguistic features that contribute to the retrieval of illocutionary force, expanding the lexical search method. It showed notable improvements over lexical search, with precision increasing to 18.63% and recall increasing to 72.20%. SVM classifier 2 represents a move toward context-aware retrieval by incorporating surface-level contextual cues into the feature-engineering process. While this improved recall to 75.00%, precision remained relatively low at 17.03%, indicating a modest improvement in capturing the targeted directives.

3.3. Baseline 3-1 and 3-2

The third retrieval experiment, based on GPT-4o, represents a further step towards a function-end approach, integrating a deeper level of contextual information to identify the directive force. We began by applying the base model, GPT-4o, to all sub-utterances in the test set via batch-level API calls, thereby establishing a baseline of 3-1. The performance is summarized in Table 6:

Table 6
Confusion matrix for the performance of GPT-4o base model on test set.

		GPT-4o	
		0	1
AN00	0	12,318 (TN)	313 (FP)
	1	198 (FN)	230 (TP)

The unmodified GPT-4o base model correctly identified 12,318 sub-utterances as non-directives, which were classified as TNs. However, 313 sub-utterances were mistakenly classified as directives, resulting in FPs. Among the manually annotated directive speech acts, 198 were not detected by the model, accounting for FNs. Despite these errors, GPT-4o successfully identified 230 directives as TPs.

The precision is 42.36%, which is notably strong given the class imbalance and is the highest among all the classifiers examined. This indicates that the GPT-4o has a substantial capacity to recognize directive forces even without fine-tuning. The recall of 53.74% shows that the model retrieved more than half of the true directives, but still missed a considerable number. The F2-score of 0.5100 reflected a moderate overall ability to accurately identify directive speech acts. At the same

time, the improvement in precision appears to be driven largely by better discrimination of non-directives. This suggests that, in its untuned form, GPT-4o may be more reliable in excluding non-directive cases.

We then fine-tuned GPT-4o on the training set, using the configuration shown in Table 7.

Table 7
Fine-tuning configuration.

Parameter	Value
Base model	GPT-4o-2024-08-06
Fine-tuned model	GPT-4o-Directives
Trained tokens	7,832,661
Training method	Supervised fine-tuning
Epochs	3
Batch size	5
LR multiplier	2
Seed	314
Training loss	0.0006

The fine-tuned model, GPT-4o-Directives, correctly identified 12,185 sub-utterances as non-directives, which were classified as TNs. However, 446 sub-utterances were incorrectly classified as directives, resulting in FPs. Additionally, 62 directive speech acts that were manually annotated were not detected by the model, representing FNs. Despite these errors, the model accurately identified 366 directives that constituted TPs. The performance is summarized in Table 8:

Table 8
Confusion matrix for the performance of fine-tuned GPT-4o model on test set.

		GPT-4o-Directives	
		0	1
AN00	0	12,185 (TN)	446 (FP)
	1	62 (FN)	366 (TP)

The precision was 45.07%, which was the highest among all methods tested. Nearly half of the predicted directives were correct despite the severe class imbalance. A recall of 85.51% represents a major improvement over previous methods and shows that the model retrieves the most true directive instances, with relatively few missed. This high recall is particularly notable, given the inherent subjectivity of directive identification and the detailed operational guidelines applied in this study. The F2-score of 0.7250 reflects a strong overall performance and demonstrates that the fine-tuned model is markedly more effective than the other retrieval approaches.

In conclusion, the integration of superficial contextual features in SVM classifier 2 led to a slight improvement in recall, although the precision remained relatively low. A more substantial enhancement was observed with the inclusion of deeper contextual information via LLM, together with the fine-tuning of the model to align with the specific annotation criteria in our study.

4. Analysis of retrieval failures

We further examined the retrieval failures for each method, including false positives (FPs) and false negatives (FNs). As the ultimate goal of our study was to cover as many directive speech acts as possible in natural conversational data, the two error types have different implications. FPs, in which non-directives are mistakenly classified as directives, are relatively minor because they add noise but do not reduce coverage. FNs, by contrast, are more consequential because they represent genuine directive speech acts that escape detection.

Table 9
Number of misclassified sub-utterances.

	False positives (FP)	False negatives (FN)
Lexical search	1196	198
SVM classifier 1	1350	119
SVM classifier 2	1564	107
GPT-4o	313	198
GPT-4o-Directives	446	62

Table 9 summarizes the numbers of FPs and FNs across the retrieval experiments. Overall, the results show a clear reduction in misclassification, particularly reflected in the decreasing number of FNs.

4.1. Failures of lexical search

In the case of lexical-search-based retrieval, 1196 sub-utterances were incorrectly classified as directives (FPs), while 198 manually identified directive speech acts were missed by this method (FNs).

4.1.1. False positives

These three main reasons account for the majority of FPs. First, performative verbs and metacommunicative expressions often appear descriptive rather than directive. While these terms for the lexical search were derived from well-established studies of speech act verbs, the mere presence of these verbs does not necessarily indicate directive force. In most cases, the directive force represents only a subset of the meanings associated with such verbs or nouns. For example:

(1) I was never allowed out till nine o'clock [S8GL, $u = 653$, $s = 2$]

Second, imperative detection in spontaneous speech was unreliable. Spoken English often omits explicit subjects, even in declarative sentences, and is characterized by frequent interruptions, fillers, and self-repairs, all of which complicate syntactic analyses. As a result, imperative-like structures lacking clear subjects or cases in which dependency parsers (i.e., spaCy) failed to identify a subject were often misclassified as directives. When dependency analysis labeled the verb as the sentence's ROOT without detecting a subject (nsubj), the subutterance was erroneously categorized as imperative.

(2) can't be getting cold and warm all the time [SN2R, $u = 186$, $s = 1$].

Third, modal and semi-modal verbs are highly context dependent. Forms such as *have to*, *should*, *can* or *might* often express descriptions or possibilities rather than obligations. In some cases, these constructions express literal or descriptive meanings, such as outlining conditions or stating requirements rather than issuing commands or influencing the hearer's actions. For example:

(3) if you don't wanna be on Facebook you should be able to properly delete your account [S2PS, $u = 702$, $s = 2$]

4.1.2. False negatives

This method also produced notable FNs. First, many imperatives were missed, because imperative detection is particularly dependent on the accuracy of dependency parsing in spaCy. Although the transformer-based en_core_web_trf model, spaCy's most accurate pipeline, was used, its performance remained constrained when applied to irregular and fragmented spoken language data. For instance:

(4) keep your hand in <anon type = "name" nameType = "m"/> you don't wanna completely retire [S55B, $u = 420$, $s = 1$]

The first half constitutes an imperative, while the second half is a separate declarative clause. Because there is no punctuation to mark these boundaries, spaCy's parser treats such case as containing multiple clauses or coordinated verbs, which prevents *keep* from being recognized as the ROOT.

Second, directives without explicit IFIDs often failed to be identified. The absence of these indicators makes such utterances more reliant on contextual and semantic cues for identification that the syntactic and lexical rules used in the retrieval process are not equipped to capture. This limitation is particularly pronounced in spontaneous spoken discourse, in which speakers frequently rely on implicit cues or less conventionalized structures to convey directive force:

(5) well your nearest public library has got all those [S7JF, $u = 305$, $s = 1$]

According to its surrounding context, the speaker in (5) suggests the hearer to visit the nearest public library so that he or she can obtain useful maps. The illocutionary force, however, is not cued by any conventionalized marker, but is conveyed through a declarative statement.

Third, multicomponent directives are difficult to detect. A complete directive speech act comprises a head act, an incomplete act, a mixed act, or an adjunct. For instance:

(6) how do you fee if you can speak up a little more [SZT4, $u = 153$, $s = 1$] because we've got some people outside [SZT4, $u = 153$, $s = 2$]

Example (6) conveys a suggestion that the addressee should speak louder due to noise from people outside. The first segment was correctly identified as a directive, as it contains the explicit pattern you can. By contrast, the second segment, which functions as an adjunct providing the rationale for the suggestion, was not recognized as part of the directive speech act.¹¹

¹¹ As illustrated here, a directive speech act may extend over more than one sub-utterance. Following Blum-Kulka and Olshtain (1984: 200), we treat directive sequences as potentially consisting of an address term, a head act, and an adjunct. Accordingly, when a directive spans multiple units, all contributing sub-utterances are annotated to capture the complete directive sequence.

4.2. Failures of SVM-based retrieval

This section presents a comparative analysis of misclassifications using SVM classifiers 1 and 2. SVM classifier 1 misclassified 1350 sub-utterances as FPs and 119 sub-utterances as FNs. SVM classifier 2 exhibited a reduction in the number of FNs, misclassifying only 107 sub-utterances. However, this improvement in reducing FNs was accompanied by an increase in FPs, with 1564 sub-utterances misclassified.

4.2.1. False positives

Most FPs in Classifier 1 stem from their overreliance on surface-level features, such as modal verbs, imperative structures, and second-person pronouns, without adequately considering deeper contextual information.

One recurring issue, already observed in lexical search, is the classifier's tendency to treat high-frequency modal verbs and directive-related keywords (*can, must, have to, should, do and go*) as strong indicators of directives.

Another common source of FPs is cases containing the second-person pronoun *you*. In conversational English, *you* is frequently used in a generalized sense, referring to an undefined audience rather than directly addressing the audience in its presence or sometimes referring to the speaker himself or herself. The classifier frequently interprets utterances as commands, failing to distinguish between the directive and non-directive uses of *you + verb* constructions:

(7) *you* can choose not to include a profile picture [S2PS, $u = 253, s = 1$].

Isolated sub-utterances often require the surrounding dialog to determine whether they constitute directives. For example, a speaker might be quoting someone else, reflecting on their own obligations, or describing hypothetical scenarios, rather than issuing directives. Without access to this broader context, the classifier struggles to accurately interpret the speaker's intent. In (9), the sub-utterance structurally resembles a directive, which may lead the model to classify it as a suggestion. However, within the context of the preceding turn, it is clear that the speaker is responding to and reinforcing a sentiment already expressed by another speaker: *'and I said to him well you do care Ben cos it's English (u = 533, s=1) and it's one of those things that you've gotta have (u = 533, s = 2).'*

(8) *you* need to get your English yeah [S8GL, $u = 534, s = 1$]

These challenges highlight the limitations of SVM classifier 1 in incorporating pragmatic and conversational nuances, emphasizing the need for a broader contextual understanding to reduce misclassification.

By incorporating superficial contextual cues such as adjacent n -grams and speaker information, Classifier 2 partially corrects these issues. It successfully corrected 486 FPs, particularly those involving self-referential statements, questions, proposals and conditional or hypothetical constructions. These include interrogatives, as well as conditional structures such as *"if ... then ..."* that serve to explore possibilities rather than to instruct. Classifier 2 also identified reported-speech constructions or references to objective rules better, where the directive-like form concerned third-party behavior rather than that of the hearer. It further corrected some fragmented cases by recognizing that incomplete forms (e.g., *"you + verb"*) were embedded within a broader descriptive sequence. However, classifier 2 also introduced 700 new FPs, underscoring the limitations of shallow contextual features in resolving form-function mismatches.

4.2.2. False negatives

Compared to lexical search, both SVM classifiers produced fewer FNs. However, classifier 1 still missed 119 directives, mainly due to its reliance on explicit lexical and syntactic markers, with many indirect and nonconventional directives remaining undetected. For example, in (17), the indirect suggestion *"it's not a bad place to eat"* expresses positive evaluation rather than containing overt directive markers.

(9) *well it's not a bad place to eat is it?* [S2LC, $u = 645, s = 1$]

In (9), the speakers are discussing where to eat during a trip. Against this background, this utterance functions as an indirect suggestion. Formally, it is a tag question, but pragmatically, it recommends the place as a possible option for eating.

Classifier 1 also struggled with incomplete or fragmented directives, the interpretations of which depended on the discourse context.

(10) *you'll have to then put that bag in to the erm [SRVR, $u = 244, s = 2$] in to the freezer [SRVR, $u = 246, s = 1$]*

In (10), the classifier identified the initial directive (*"you'll have to ..."*) but failed to capture the follow-up clause functioning as part of the same directive.

Classifier 2 shows a moderate improvement in handling indirect and fragmented directives as well as multi-turn expressions. The addition of surface-level contextual cues allowed it to detect subtle directive intent more effectively in cases where classifier 1 relied too heavily on explicit markers. Classifier 2's improved sensitivity is reflected in its reduced FN count (107). However, while these enhancements mark progress, they do not fully address several persistent challenges. Classifier 2 struggles with highly indirect or ambiguous expressions that require deep pragmatic inferences to uncover directive intent. Fragmented utterances and multiturn dependencies remain problematic, particularly when critical information is scattered across turns. Furthermore, the model continues to inconsistently identify peripheral adjuncts, particularly when their directive relevance hinges on subtle contextual or rhetorical cues. Although Classifier 2 demonstrated clear progress, further refinement is required, particularly through the incorporation of deeper contextual information.

Overall, a comparison between classifier 1 and classifier 2 shows that incorporating superficial contextual information yields measurable but incomplete improvements. While classifier 2 reduces some types of FNs and corrects certain categories of FPs, its increased FP rate demonstrates that shallow contextual features alone are insufficient for fully resolving the ambiguity inherent in spoken context-dependent directive speech acts. More robust modeling of the discourse context and pragmatic inference is required to substantially enhance the retrieval accuracy.

4.3. Failures of LLM-based retrieval

GPT-4o generated 313 FPs and 198 FNs. Its FP rate was markedly lower than that of the lexical search and both SVM classifiers. Fine-tuning substantially reduced the number of FNs (62) but increased the number of FPs to 446.

4.3.1. False positives

The most noticeable improvement over previous retrieval methods is that the GPT-4o base model no longer indiscriminately classifies all sub-utterances containing modal verbs potentially associated with directive force as directive speech acts. It can be said that the false positives (FPs) introduced by GPT-4o are “more sophisticated errors,” which, to some extent, resemble the mistakes made by human annotators.

The first type of FPs consists of borderline cases, which are inherently controversial. For instance, the model sometimes struggles to distinguish between commissives and directives, a difficulty also faced by human annotators. In this study, we adopt an operational criterion based on responsibility for the action: the speaker in commissives and the hearer in directives. Even so, this distinction is not always clear-cut in natural dialog. Example (11) illustrates such ambiguity:

(11) can I offer you a half of this [S55B, $u = 65$, $s = 1$]

While *offer* is typically associated with commissive force as a performative verb (e.g., *I offer to help you*), the utterance functions in context as an implicit request. The speaker appears to be inviting the hearer to take some of the food in order to reduce the speaker's own share. In this sense, the utterance functions pragmatically as a request, despite its commissive form.

The second error type arises from the operationalized annotation guidelines established in this study. Due to practical constraints, the prompts provided to GPT-4o did not encode the full set of fine-grained annotation criteria, but instead relied on a general definition, scope, and examples of directive speech acts. As a result, the model occasionally failed to match the level of precision specified in the guidelines. One such case concerns responsive directives in reported speech, which are inherently directive but fall outside the scope of the present study, and were not consistently excluded by the model, as illustrated in (12) and (13).

(12) yes please [SRBZ, $u = 388$, $s = 1$]

(13) I said you're allowed a bit of that [S8GL, $u = 255$, $s = 3$]

(12) Is the response to the commissive of offering milk ‘*you said yes to milk yeah?*’ (SRBZ, $u = 254$, $s = 1$) is classified as a responsive directive. (13) represents a directive in reported speech. Both are intentionally excluded from the study's scope, according to our annotation guidelines (see Zhou and Xu, *in press*).

Notably, the fine-tuned model, GPT-4o-Directives, systematically improved upon the FPs, particularly those associated with annotation guidelines as reported in base model. Of the 313 FPs generated by GPT-4o, GPT-4o-Directives successfully corrected 229. However, when considering the total number of false positives produced by the GPT-4o-Directives, there was an increase in the error rate; the fine-tuned model resulted in 446 false positives. In other words, while this fine-tuned model showed better alignment with the annotation guidelines in some cases, it also introduced additional misclassifications.

(14) and then we if we have dinner about eight or something [S2LC, $u = 194$, $s = 1$] that that's alright isn't it? [S2LC, $u = 194$, $s = 2$] it's not too late [S2LC, $u = 194$, $s = 2$]

(15) it's not getting too late in the day to be eating at that time is it? [S2LC, $u = 196$, $s = 2$]

One prominent source of these new FPs appears to be the overgeneralization of the annotation guidelines, particularly in cases involving adjuncts. For example, in (14), the fine-tuned model successfully identified the sequence as a directive speech act. In (15), however, the model also classified the sub-utterance as directive. Human annotators did not treat this case as an adjunct, since it is separated from the sequence in (14) by another turn. (15) is itself a somewhat borderline case, the utterance occurs in the context of a road trip and concerns when to stop for a meal. It may therefore be interpreted as an indirect suggestion to stop and eat at a particular time, but it may also be understood more weakly as part of a discussion about the timing of the meal. This illustrates a tendency of the fine-tuned model to extend directive interpretations too broadly across discourse and thus to overgenerate directive classifications in contexts that are only loosely connected to an earlier directive sequence.

4.3.2. False negatives

Although the unmodified GPT-4o base model demonstrated improved performance in terms of precision, its recall of directive speech acts was comparatively modest, yielding results on par with those of lexical search-based retrieval.

We adopted a relatively broad definition of directive speech acts, which included borderline cases not typically examined in traditional speech act research. However, the prompts provided to GPT-4o did not fully reflect the detailed specifications of each annotation guideline. As a result, the model failed to recognize these cases. Examples include:

(16) imagine if we went back [SY8A, $u = 173$, $s = 1$]

(17) so hang on you get up at half past and you leave at seven forty-seven [SRVR, $u = 24$, $s = 2$]

The base model's inability to recognize these cases highlights its limitations in identifying nonstandard directive forms, particularly those that are highly conventionalized or have undergone partial grammaticalization.

Second, the annotation guidelines for this study required a complete directive speech act. A given sub-utterance may be annotated as a head act, an incomplete fragment of the head act, an adjunct, or a combination of a head act and adjuncts. However, owing to the constraints of the prompts provided, the unmodified GPT-4o base model demonstrates limited capacity, identifying only the core speech act itself while frequently neglecting related components.

Finally, although GPT-4o has demonstrated some contextual capabilities, it remains prone to misclassification in cases that are more implicit or subjective. It is essential to recognize that directive speech acts, as functionally defined units, inherently involve subjectivity. Even human annotators often disagree on whether a particular sub-utterance has directive force. Thus, the model's misjudgments in this regard are not entirely unexpected.

(18) erm does such a thing exist as a digital little really tiny alarm clock? [SRBZ, $u = 22$, $s = 1$] they must mustn't they? [SRBZ, $u = 22$, $s = 2$]

In (18), the speaker appears to be indirectly suggesting that the hearer might consider purchasing a '*digital little really tiny alarm clock*.' This interpretation is informed by the context in which the hearer has previously expressed dissatisfaction with the noise of traditional alarm clocks, leading the speaker to propose a possible solution, namely acquiring a smaller digital clock. However, this interpretation carries a degree of subjectivity, as the speaker may merely be inquiring about the existence of such an item rather than making a suggestion.

The fine-tuned model, GPT-4o-Directives, demonstrated a significant reduction in FNs (62). This improvement is particularly noteworthy, given the primary objective of these retrieval experiments, which was to maximize the identification of directive speech acts. By ensuring that as many positive instances as possible are accurately retrieved in an imbalanced dataset, the model effectively addresses one of the critical challenges in this task.

This improvement reflects the fine-tuned model's closer alignment with our annotation guidelines, enabling it to capture non-typical and context-dependent indirect directives that the base GPT-4o model failed to identify, including cases such as (5).

5. Results and discussion

This study conducted retrieval experiments employing three sets of distinct automated methods and established performance baselines by measuring precision, recall, and F2-score. These retrieval methods can be situated along a continuum ranging from form-dependent to function-oriented approaches.

Table 10

Performance of different methods on the directive retrieval task.

	Retrieval method	Precision	Recall	F2-score
Baseline 1	Lexical search	16.13%	53.74%	0.3665
Baseline 2-1	SVM classifier 1 (<i>lexical, syntactical, semantic features</i>)	18.63%	72.20%	0.4583
Baseline 2-2	SVM classifier 2 (<i>lexical, syntactical, semantic and superficial contextual features</i>)	17.03%	75.00%	0.4462
Baseline 3-1	GPT-4o base model	42.36%	53.74%	0.5100
Baseline 3-2	The fine-tuned model GPT-4o-Directives	45.07%	85.51%	0.7250

The results in Table 10 reveal the limitations of purely lexicogrammatical approaches and highlight the potential of deep contextual modeling in speech act retrieval. Traditional rule-based (Baseline 1)¹² and feature engineering (Baseline 2-1) methods often fall short when faced with indirectness without explicit markers, particularly those requiring pragmatic inferences or context-dependent interpretations. Although the addition of superficial contextual features yields some improvement, the overall gain remains modest (Baseline 2-2). Instead of relying on discrete rule sets or predefined mappings, LLMs dynamically construct meaning representations by processing language inputs as high-dimensional vectors contextualized within vast distributions of prior linguistic experience. This allows them to go beyond isolated lexical, syntactic, or even semantic features and instead holistically attend to the pragmatic context in which an utterance occurs. As a result, they are better able to identify directive speech acts even when no overt directive marker is present and the relevant

¹² It is worth noting, however, that lexical search and the general-purpose GPT-4o model achieved the same recall in the present dataset, despite substantial differences in precision and overall performance. This suggests that even traditional methods could retrieve a meaningful proportion of relevant cases, which helps explain their methodological usefulness in pre-LLM corpus research.

illocutionary force must be inferred. The comparison between Baseline 2-2 and Baseline 3-2 is particularly revealing in this respect. Both models incorporate contextual information, but the fine-tuned GPT-4o model performs substantially better than the SVM classifier with superficial contextual features. This suggests, first, that the mere inclusion of context is not in itself sufficient: what matters is how context is represented and integrated into the retrieval process. Second, it indicates that surface-level contextual features capture only a limited part of the information relevant to directive interpretation. Third, the contrast points to the importance of richer sequential and interactional context, including what precedes and follows the target utterance, for the identification of directive force. In short, the comparison between Baseline 2-2 and Baseline 3-2 suggests that successful directive retrieval depends less on the presence of contextual information as such than on the model's ability to represent and exploit context in pragmatically relevant ways.

Thus, LLM-based retrieval constitutes a “holistic approach” (or a “third route”) to speech act identification, complementing and extending the form-to-function (Aijmer, 1996; Deutschmann, 2003; Jucker et al., 2008; Jucker and Taavitsainen, 2014 among others) and function-to-form (Garcia, 2004; Milà-Garcia, 2018; Flöck and Geluykens, 2018 among others) paradigms. Unlike canonical corpus search, which relies on presupposed mappings between linguistic forms and speech functions or requires human annotators to interpret utterances line by line, LLM unifies form and function not through explicit lexical-grammatical rule mapping but via emergent, embedding-based distributional pattern computation. It addresses the key challenges in corpus-based research on speech acts and paves the way for future studies on computational pragmatics.

6. Conclusion and future directions

Overall, this study has explored a longstanding methodological challenge in pragmatics – the retrieval of speech acts from naturally occurring conversation. Specifically, we focus on directives, a category characterized by form-function mismatch. Through three retrieval experiments, we showcased that a context-enriched LLM-based approach outperformed methods relying on lexical, syntactic, semantic, and superficial contextual features. We also examine the retrieval failures of the different strategies, using them as a window into how the respective methods operate.

These findings are particularly encouraging for corpus-based pragmatic research on speech acts and contribute to a growing body of work exploring the use of LLM-based methods in pragmatics (see Yu et al., 2024 among others). While speech act theory, as originally formulated and developed in the works of Austin and Searle, relied heavily on theoretical reflection and constructed examples, corpus pragmatics has opened up new empirical perspectives by examining speech acts in naturally occurring language data. Yet empirical work on naturally occurring interaction has remained comparatively difficult, not least because relevant speech act instances are hard to retrieve in a systematic and scalable way. The present study suggests that an LLM-based approach can substantially improve retrieval efficiency and may therefore help make larger-scale corpus-pragmatic research on authentic interaction more feasible. It may create new opportunities to complement established, largely top-down theoretical accounts with bottom-up empirical evidence derived from naturally occurring data.

The present study suggests that LLM-based retrieval may also be applicable to other speech act categories. This is especially promising because directives constitute a particularly difficult test case for automated retrieval, given the pronounced form-function mismatch. A natural next step, therefore, is to investigate how far this approach generalizes across speech act types. Another important question concerns the extent to which LLM performance depends on factors such as training strategy and the nature of the input examples. A systematic comparison of model variants in this respect would be valuable. In addition, the present study examined only two LLM-based strategies, prompting and fine-tuning. Further work should therefore compare a wider range of approaches, including retrieval-augmented generation (RAG) and other parameter-efficient methods, with a view to identifying LLM-based retrieval solutions that are both effective and methodologically economical.

Some caution is nevertheless needed in interpreting retrieval failures. Although this study examined false positives and false negatives through selected examples, such interpretations inevitably remain subjective to some extent. Future research should revisit these issues on a broader empirical basis, using larger and more balanced corpora to examine whether recurrent patterns of retrieval failure can be identified more robustly. Such work would not only refine the evaluation of retrieval methods, but also deepen our understanding of the points at which pragmatic meaning continues to resist formalization.

CRedit authorship contribution statement

Guying Zhou: Writing – original draft, Visualization, Methodology, Data curation, Conceptualization. **Jiajin Xu:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Funding

This work was supported by the China MOE Major Project Fund of Key Research Center for Humanities and Social Sciences, 22JJD740012.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

Prompt used for batch-calling process of GPT-4o:

system=

""

You are an expert in annotating speech acts in spoken English. Your task is to identify directive speech acts in naturally occurring English conversations.

A directive speech act is one in which the speaker attempts to get the hearer to perform an action. These utterances express the speaker's wishes, desires, or commands, and are intended to influence the behavior of the hearer. Examples of directive speech acts include requests, commands, suggestions, invitations, and advice. However, do not include interrogatives whose function is to request information.

Your annotation will focus specifically on action-directives as described above.

Instructions for Annotation:

- 1 You will be given one target sub-utterance, along with several sub-utterances before and after it as context. Your task is to judge the target utterance based on the provided context.
- 2 Annotate the target sub-utterance in the provided XML text: If the sub-utterance is a directive speech act, label it as positive (1). If the sub-utterance is not a directive speech act, label it as negative (0).
- 3 Carefully read the context surrounding the target utterance, make sure your judgment is contextually informed.

""

user_tmp=

""

Context:

{context}

Target Sub-utterance:

{target}

Please only return the label (1 or 0).

""

Data availability

Data will be made available on request.

References

- Aijmer, K., 1996. *Conversational Routines in English: Convention and Creativity*. Longman.
- Arizavi, S., Choubasaz, Y., 2019. To use or not to use the shorter forms: a corpus-based analysis of the apologetic expressions "sorry and i'm sorry" in American spoken English discourse. *Corpus Pragmat.* 3 (1), 21–47.
- Austin, J.L., 1962. *How to Do Things with Words*. Clarendon Press.
- Bach, K., Harnish, R.M., 1979. *Linguistic Communication and Speech Acts*. MIT Press.
- Bell, L.B., 2020. Illocution on Twitter: the Construction and Analysis of a Social Media Speech Act Corpus. Georgetown University (Ph.D. dissertation).
- Bilbow, G.T., 2002. Commissive speech act use in intercultural business meetings. *Int. Rev. Appl. Linguist. Lang. Teach.* 40 (4), 287–303.
- Blum-Kulka, S., Olshtain, E., 1984. Requests and apologies: a cross-cultural study of speech act realization patterns (CCSARP). *Appl. Linguist.* 5 (3), 196–213.
- Bunt, H., 2022. The multifunctionality of utterances in interactive discourse. In: Yin, Z., Vine, E. (Eds.), *Multifunctionality in English: Corpora, Language and Academic Literacy Pedagogy*. Routledge, pp. 11–29.
- Burnette, J., Calude, A.S., 2022. Wake up New Zealand! directives, politeness and stance in Twitter #Covid19NZ posts. *J. Pragmat.* 196, 6–23.
- Cortes, C., Vapnik, V., 1995. Support-vector networks. *Mach. Learn.* 20 (3), 273–297.
- Degenhardt, J., Bernaisch, T., 2022. Apologies in South Asian varieties of English: a corpus-based study on Indian and Sri Lankan English. *Corpus Pragmat.* 1–23.
- Deutschmann, M., 2003. *Apologising in British English*. Umeå University Electronic Publishing (Doctoral dissertation, Umeå University).
- Deutschmann, M., 2006. Social variation in the use of apology formulae in the British national corpus. In: Renouf, A., Kehoe, A. (Eds.), *The Changing Face of Corpus Linguistics*. Rodopi, Amsterdam/New York, pp. 205–221.
- Duran, N., Battle, S., Smith, J., 2023. Sentence encoding for dialogue act classification. *Nat. Lang. Eng.* 29 (3), 794–823.
- Gagne, C., 2018. Indirectness and entitlement in product requests in British service encounters. *J. Pragmat.* 133, 1–14.
- Garcia, P., 2004. *Meaning in Academic Contexts: a Corpus-based Study of Pragmatic Utterances*. Northern Arizona University (Ph.D. dissertation).
- Honnibal, M., Montani, I., 2017. spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing [Software]. Explosion AI. <https://spacy.io>.
- John, P., Brooks, B., Schriever, U., 2019. Speech acts in professional maritime discourse: a pragmatic risk analysis of bridge team communication directives and commissives in full-mission simulation. *J. Pragmat.* 140, 12–21.
- Jucker, A.H., 2009. Speech act research between armchair, field and laboratory: the case of compliments. *J. Pragmat.* 41 (8), 1611–1635.
- Jucker, A.H., 2018. Apologies in the history of English: evidence from the Corpus of Historical American English (COHA). *Corpus Pragmat.* 2 (4), 375–398.

- Jucker, A.H., Taavitsainen, I., 2014. Complimenting in the history of American English. In: Taavitsainen, I., Jucker, A.H., Tuominen, J. (Eds.), *Diachronic Corpus Pragmatics*. John Benjamins, Amsterdam, pp. 257–276.
- Jucker, A.H., Schneider, G., Taavitsainen, I., Breustedt, B., 2008. Fishing for compliments precision and recall in corpus-linguistic. In: Jucker, A.H., Taavitsainen, I. (Eds.), *Speech Acts in the History of English*. John Benjamins, Amsterdam/Philadelphia, pp. 273–294.
- Jucker, A.H., 2017. Speech Acts and Speech Act Sequence: Greetings and Farewells in the History of American English. *Stud. Neophilol.* 89 (sup1), 39–58.
- Jurafsky, D., Martin, J.H., 2024. Speech and language processing: an introduction to natural language processing. In: *Computational Linguistics, and Speech Recognition*, third ed. draft of February 3, 2024). <https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf>
- Keizer, S., 2001. A Bayesian approach to dialogue act classification. In: Kühnlein, P., Rieser, H., Zeevat, H. (Eds.), *BI-DIALOG 2001: Proceedings of the 5th Workshop on Formal Semantics and Pragmatics of Dialogue*, pp. 210–218.
- Kitao, S.K., 2012. Using a spoken corpus compiled from subtitles to study apologies. *Asphodel* 47, 50–77.
- Kitao, S.K., Kitao, K., 2012. Using spoken corpora to study request forms and strategies. In: *Presented at Eurocall 2012 (Gothenburg)*.
- Koester, A.J., 2002. The performance of speech acts in workplace conversations and the teaching of communicative functions. *System* 30 (2), 167–184.
- Kohnen, T., 2008. Directives in old English. In: Jucker, A.H., Taavitsainen, I. (Eds.), *Speech Acts in the History of English*. John Benjamins Publishing Company, pp. 27–44.
- Kohnen, T., 2015. Speech acts: a diachronic perspective. In: Aijmer, K., Rühlemann, C. (Eds.), *Corpus Pragmatics: a Handbook*. Cambridge University Press, pp. 52–83.
- Kumar, H., Agarwal, A., Dasgupta, R., Joshi, S., 2018. Dialogue act sequence labeling using hierarchical encoder with CRF. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*. New Orleans, LA, pp. 3440–3447.
- Lee, J.Y., Dernoncourt, F., 2016. Sequential short-text classification with recurrent and convolutional neural networks. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 515–520 (San Diego, CA).
- Levinson, S.C., 2017. Speech acts. In: Huang, Y. (Ed.), *The Oxford Handbook of Pragmatics*. Oxford University Press, pp. 199–216.
- Louwerse, M., Crossley, S., 2006. Dialog act classification using n-gram algorithms. In: Sutcliffe, G., Goebel, R. (Eds.), *Proceedings of the 19th International Florida Artificial Intelligence Research Society Conference*, pp. 758–763 (Melbourne Beach, FL).
- Love, R., Dembry, C., Hardie, A., Brezina, V., McEnery, T., 2017. The spoken BNC2014: designing and building a spoken corpus of everyday conversations. *Int. J. Corpus Linguist.* 22 (3), 319–344.
- Lutzky, U., Kehoe, A., 2017. “I apologise for my poor blogging”: searching for apologies in the Birmingham Blog Corpus. *Corpus Pragmat.* 1 (1), 37–56.
- Merrison, A.J., Wilson, J.J., Davies, B.L., Haugh, M., 2012. Getting stuff done: comparing e-mail requests from students in higher education in Britain and Australia. *J. Pragmat.* 44 (9), 1077–1098.
- Milà-García, A., 2018. Pragmatic annotation for a multi-layered analysis of speech acts: a methodological proposal. *Corpus Pragmat.* 2 (3), 265–287.
- Nagy, C.K., Németh, T., Enikő & Németh, Z., 2025. How to investigate implicit pragmatic phenomena in corpora. *J. Pragmat.* 240, 79–90.
- Ortega, D., Vu, N.T., 2017. Neural-based context representation learning for dialog act classification. In: *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 247–252 (Saarbrücken, Germany).
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al., 2011. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.
- Qadir, A., Riloff, E., 2011. Classifying sentences as speech acts in message board posts. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 748–758 (Edinburgh).
- Raheja, V., Tetreault, J., 2019. Dialogue act classification with context-aware self-attention. *arXiv preprint arXiv:1904.02594*.
- Ribeiro, E., Ribeiro, R., de Matos, D.M., 2015. The influence of context on dialogue act recognition. *arXiv preprint arXiv:1506.00839*.
- Ritter, A., Cherry, C., Dolan, B., 2010. Unsupervised modeling of Twitter conversations. In: *Human Language Technologies: the 2010 Annual Conference of the North American Chapter of the ACL*, pp. 172–180.
- Rühlemann, C., 2010. What can a corpus tell us about pragmatics? In: O’Keeffe, A., McCarthy, M.J. (Eds.), *The Routledge Handbook of Corpus Linguistics*. Routledge, pp. 288–301.
- Rühlemann, C., 2019. *Corpus Linguistics for Pragmatics: a Guide for Research*. Routledge.
- Schauer, G.A., Adolphs, S., 2006. Expressions of Gratitude in Corpus and DCT Data: Vocabulary, Formulaic Sequences, and Pedagogy. *System* 34 (1), 119–134.
- Searle, J.R., 1969. *Speech Acts: an Essay in the Philosophy of Language*. Cambridge University Press.
- Searle, J.R., 1979. *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge University Press.
- Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C., Meteer, M., 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Comput. Linguist.* 26 (3), 339–373.
- Su, H., 2017. Local grammars of speech acts: an exploratory study. *J. Pragmat.* 111, 72–83.
- Surendran, D., Levow, G.A., 2006. Dialog act tagging with support vector machines and hidden Markov models. In: *Presented at Interspeech 2006 and 9th International Conference on Spoken Language Processing*, pp. 1950–1953. Pittsburgh, USA.
- Vapnik, V., 1998. *Statistical Learning Theory*. Wiley.
- Wierzbicka, A., 1987. *English Speech Act Verbs: a Semantic Dictionary*. Academic Press.
- Wong, M.L.Y., 2010. Expressions of Gratitude by Hong Kong Speakers of English: Research from the International Corpus of English in Hong Kong (ICE-HK). *J. Pragmat.* 42 (5), 1243–1257.
- Yu, D., Li, L., Su, H., Fuoli, M., 2024. Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: the case of apology. *Int. J. Corpus Linguist.* 29 (4), 534–561.
- Zhang, R., Gao, D., Li, W., 2011. What are tweeters doing: recognizing speech acts in Twitter. In: *Presented at Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence (San Francisco, USA)*.
- Zhou, G., Xu, J., n.d.. (in press). dassBNC2014: Retrieving and mapping directive speech acts in an open-domain corpus of everyday conversations. *Int. J. Corpus Linguist.*
- Zhou, G., 2025. *A Corpus-Based Study of Variation in Indirectness Across English Directive Speech Acts*. Ph.D. dissertation. Beijing: Beijing Foreign Studies University.

Guying Zhou is lecturer at the School of German Studies, Beijing Foreign Studies University. Her research interests include corpus linguistics, corpus pragmatics, contrastive linguistics, and German linguistics.

Jiajin Xu is professor of linguistics at the National Research Centre for Foreign Language Education, Beijing Foreign Studies University, and president of the Corpus Linguistics Society of China. His research interests include corpus linguistics, discourse studies, second language acquisition, and contrastive linguistics and translation studies. He has published extensively in leading journals (e.g., *Corpus Linguistics and Linguistic Theory*, *Corpus Pragmatics*, *Discourse and Society*, *International Journal of Corpus Linguistics*, *Journal of Historical Pragmatics*, *Journal of Pragmatics*, *Language & Communication*, *Semiotica* and *Text & Talk*).