Toward Proactive Dialogic AI Agents

Sofia Brenna[1,2,*], Elisabetta Jezek[3] and Bernardo Magnini[2]

[1] *Free University of Bozen-Bolzano, 3 Dominikanerplatz - Piazza Domenicani 3, Bozen-Bolzano (BZ), 39100, Italy*
[2] *Fondazione Bruno Kessler, Via Sommarive 18, Povo, Trento (TN), 38123, Italy*
[3] *Università degli Studi di Pavia, Corso Strada Nuova 65, Pavia (PV) 27100, Italy*

Abstract

This paper introduces an ongoing research on the development of a proactive dialogic AI agent, focusing on enhancing an LLM's pragmatic competence in goal-oriented dialogues. We investigate proactivity as a collaborative behaviour that enables to provide relevant and useful information that has not been explicitly requested, thereby improving interaction efficiency and dialogue naturalness. Our approach is grounded in a corpus-based analysis of proactive behaviours in human-human dialogues across five goal-oriented dialogue corpora, leading to the creation of the D-Pro Corpus, a manually annotated resource for studying proactivity. Its analysis provides information on qualitative and distributional features of proactivity in human dialogues, as well as clues on recurrent linguistics structures that co-occur with the display of proactive behaviours. We then leverage the D-Pro Corpus to evaluate the performance of a GPT-4o model in proactivity annotation, addressing the task by providing a 4-turns context size and by targeting the last utterance for proactivity prediction. By experimenting with parameter setting and prompt configurations, we assess the model's performance across multiple dialogue corpora, obtaining encouraging results toward human-like performance, particularly with the NESPOLE! corpus. We propose to advance our research by exploring the potential of open-source models for cost-effective, large-scale automatic annotation of unlabelled dialogic data. As a final step we plan to use the large-scale annotated corpus to instruction-tune an open model, expanding its pragmatic competence for the development of more proactive and contextually aware dialogic AI system and more natural human-machine conversation.

Keywords

goal-oriented dialogues, pragmatics, proactivity, automated annotation, large language models

CEUR-WS.org/Vol-3957/BEHAVEAI-paper02.pdf

1. Introduction

One of the main interests in Artificial Intelligence regards *dialogue* and human interactions, as AI advancements have placed interest into designing machines capable of mimicking linguistic and conversational abilities in natural language. While *dialogue systems* have been investigated for a long time, attention has been mainly given to their effectiveness, namely, the capacity to achieve a communicative goal. However, it is crucial to investigate also *how* such dialogue goals are achieved through pragmatic strategies such as *collaborative behaviours*.

Human dialogue is a complex interaction characterised by systematic, coordinated behaviours and a collaborative effort on the part of each participant to communicate [1, 2]. There is extreme variability on many levels in human dialogues, depending on the complexity and the nature of the communication context. Dialogues vary widely in terms of *participants*, *initiative*, *means*, and *purposes* of interaction, leading to their classification into different categories. One common approach is to categorize dialogues based on their purpose. For example, some dialogues are goal-oriented, with participants communicating to achieve specific objectives, while others involve information seeking, argumentation, explanation, instruction giving, chit-chat, or recommendations. In the development of dialogue, the conversational competence [3] of participants plays a crucial role, as it encompasses the pragmatic skills necessary for successful communication and the ability to determine *if, when*, and *how* to use collaborative strategies [4] in order to facilitate effective and seamless interaction.

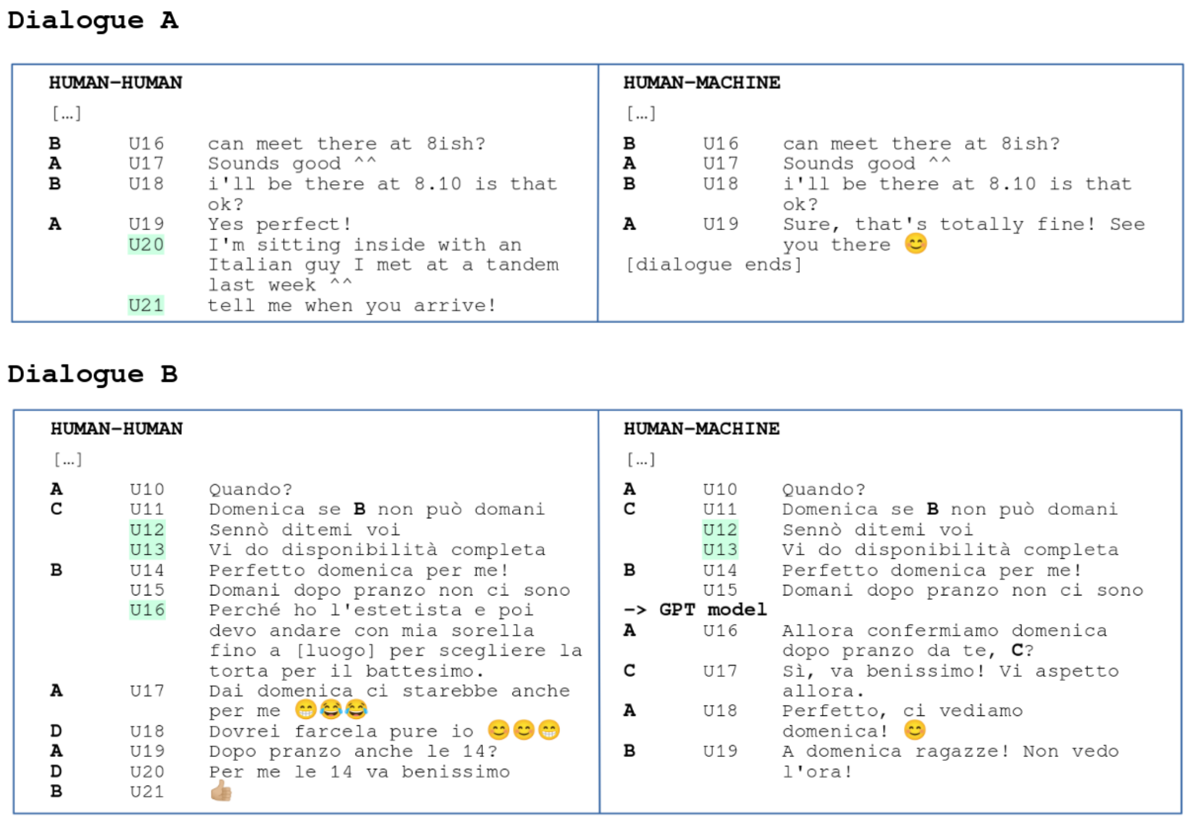
Joint Proceedings of the ACM IUI Workshops 2025, March 24-27, 2025, Cagliari, Italy

[*] Corresponding author.

sbrenna@fbk.eu (S. Brenna); elisabetta.jezek@unipv.it (E. Jezek); magnini@fbk.eu (B. Magnini)
0009-0001-3748-1448 (S. Brenna); 0000-0003-2518-5200 (E. Jezek); 0000-0002-0740-5778 (B. Magnini)

© 2025 Copyright © 2025 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073



Dialogue A

HUMAN-HUMAN

[…]

B	U16	can meet there at 8ish?
A	U17	Sounds good ^^
B	U18	i'll be there at 8.10 is that ok?
A	U19	Yes perfect!
	U20	I'm sitting inside with an Italian guy I met at a tandem last week ^^
	U21	tell me when you arrive!

HUMAN-MACHINE

[…]

B	U16	can meet there at 8ish?
A	U17	Sounds good ^^
B	U18	i'll be there at 8.10 is that ok?
A	U19	Sure, that's totally fine! See you there 😊

[dialogue ends]

Dialogue B

HUMAN-HUMAN

[…]

A	U10	Quando?
C	U11	Domenica se **B** non può domani
	U12	Sennò ditemi voi
	U13	Vi do disponibilità completa
B	U14	Perfetto domenica per me!
	U15	Domani dopo pranzo non ci sono
	U16	Perché ho l'estetista e poi devo andare con mia sorella fino a [luogo] per scegliere la torta per il battesimo.
A	U17	Dai domenica ci starebbe anche per me 😁😂😂
D	U18	Dovrei farcela pure io 😊😊😁
A	U19	Dopo pranzo anche le 14?
D	U20	Per me le 14 va benissimo
B	U21	👍🏼

HUMAN-MACHINE

[…]

A	U10	Quando?
C	U11	Domenica se **B** non può domani
	U12	Sennò ditemi voi
	U13	Vi do disponibilità completa
B	U14	Perfetto domenica per me!
	U15	Domani dopo pranzo non ci sono
-> GPT model		
A	U16	Allora confermiamo domenica dopo pranzo da te, **C**?
C	U17	Sì, va benissimo! Vi aspetto allora.
A	U18	Perfetto, ci vediamo domenica! 😊
B	U19	A domenica ragazze! Non vedo l'ora!

Figure 1: Comparison of human-human dialogue and corresponding human-machine dialogue: the full dialogue context is given to the model up to the turn shift at U18 in **A**, and to U15 in **B**, and the model is supposed to output the following utterances. Proactive utterances are highlighted in green. Both dialogue excerpts are taken from the Italian Whatsapp Corpus [5].

With the term *collaborative behaviours* in dialogue we refer to the various actions and strategies employed by participants to work together towards effective communication, shared understanding, and achieving conversation goals. These behaviours help maintain the flow, coherence, and relevance of the dialogue while ensuring that all participants have the opportunity to contribute and be heard. The concept of collaborative behaviours in dialogue does not stem from a single theory but rather arises from the integration of various theories and models in linguistics and communication studies. Among them we mention H. Paul Grice's *cooperative principle* and *maxims of conversation* [6, 7], the *speech act theory* by J. L. Austin [8] and the following works by D. Traum [9] and H. Bunt [10, 11, 12], the notion of *face* [13, 14, 15], *politeness* [13, 15], and *speech accommodation theory* [16, 17, 18] and *communication accommodation theory* [19]. Collaborative behaviours encompass a range of linguistic and pragmatic strategies enacted at different linguistic levels in dialogue, and it is known that there is a fair number of linguistic expedients, or techniques, that participants can use for collaborative purposes in a dialogue, namely, giving examples, grounding [20, 21, 22, 23, 24, 25], clarification requests [26, 27, 28, 29, 30], backchanneling [31, 32, 33], reformulation [34], convergence/divergence/maintenance [35], and proactivity [4, 36, 37, 38, 39, 40, 40].

In our research, we place particular emphasis on *proactivity* as a central aspect of collaborative dialogues; it is regarded as the ability to provide the addressee with some useful and not explicitly requested information. Works on proactivity address turn-taking strategies [41], where a proactive system takes initiative in multi-party conversations rather than waiting to respond reactively. Other works on proactivity mostly relate to the field of human-computer interaction [40, 42, 3, 43]. An example is the ProDial corpus [44], a proactive human-machine dialogue corpus, where proactive behaviour was modelled through a serious game in which an autonomous assistant employed four proactive actions—none, notification, suggestion, and intervention—to serve as the user's personal advisor in a

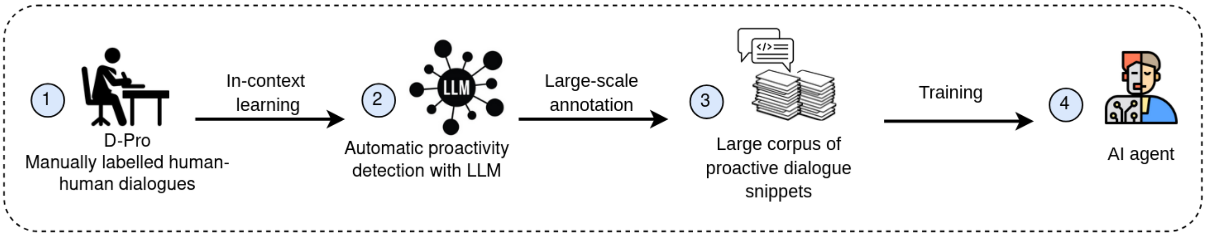

1
D-Pro
Manually labelled human-human dialogues
In-context learning
2
LLM
Automatic proactivity detection with LLM
Large-scale annotation
3
Large corpus of proactive dialogue snippets
Training
4
AI agent

Figure 2: The four phases in our research, from a corpus-based human dialogue analysis, through the automatic detection and annotation of proactivity, to the training of a proactive agent.

sequential planning task. The corpus serves as a valuable resource for understanding how proactivity influences user trust and enhances interaction efficiency [45, 46]. Another contribution is brought by Liao et al. [37], who introduce and discuss methods—including Reinforcement Learning—to equip agents with the ability to interact with end users in a proactive way.

2. Challenges

Current Large Language Models (LLMs) often do not interact collaboratively according to the Gricean cooperative principle's maxims. Despite significant advancements in the latest language models, especially models trained with instruction tuning techniques, some challenges still remain in mirroring the whole conversational competence of a human being, such as in the ability to maintain coherent and contextually relevant conversations over extended periods of time and to precisely and systematically enact human-like collaborative behaviours according to the maxims of quality, quantity, relation, and manner. We find that language models usually address the interaction with the user in an encyclopedic approach, and often interpret user interactions as task-oriented requests, adopting an omniscient, authoritative role. This omniscient attitude prevents collaboration and human-like communication even in models whose declared main feature is being *conversational* models, as for the ChatGPT interface. As much as some more conversational attitude can be induced into the model by prompting it with instructions to enhance friendliness and personification, the interaction is lacking typical empathy-driven and collaborative human-like behaviours [47]. For instance, in Figure 1, Dialogue A, the comparison between a collaborative human-human dialogue and its human-machine counterpart shows that, where humans produce proactive information and requests (U20, U21), the model ends the conversation without any proactive behaviour. In Dialogue B, where humans produce a proactive motivation for their negative response, the model concludes the turn without further collaboration. Even in cases where the model's output contains proactive content, we find that it still lacks the proper competence of *if, when*, and *how*, while it should balance behaviours without excessive "info dumping" or verbosity.

3. Methodology

To address these concerns, we articulate our research into four phases (Figure 2). The first phase involves studying the distributional and qualitative aspects of proactivity in actual dialogues, by focusing on human-human interactions from five different goal-oriented dialogue corpora. We develop an annotation schema for the utterance-level labelling of proactivity and dialogue acts, investigating: (i) how many proactive utterances are present in dialogues; (ii) how proactive utterances typically relate to preceding utterances; (iii) through which dialogue acts proactivity manifests; (iv) whether proactivity plays a role in recovering from goal-failure situations; (v) where in the dialogue is proactivity located. This leads to the creation of the manually annotated resource called *D-Pro Corpus* and provides insights for theoretical research on proactivity. The second phase consist in employing our proactivity-oriented resource, D-Pro Corpus, to assess a LLM's ability to annotate proactivity, based on previous works on automatic labelling with LLMs [48, 49, 47]. We investigate GPT-4o's ability to produce context-aware annotations in order to label proactivity, given an unlabelled dialogue in a few-shot approach. In particular, we refine the annotation task to classifying the final utterance of a 4-turn dialogue snippet as either ”proactive” or ”not proactive”. Using the D-Pro Corpus as ground truth, we evaluate various prompt configurations and the model's performance against human labels, assessing accuracy and inter-annotator agreement between gold human labels and LLM's labels. Having validated GPT-4o as an effective proactivity annotator, the next step in our research involves testing open-source models, such as Llama3 [50], for the same task. This allows for fine-tuning experiments to determine if a cheaper yet equally valid automatic proactivity annotator can be achieved. Such a model would enable quick and efficient identification of proactive utterances in a large unlabelled dialogue corpora. Further testing across different domains is possible with minimal manual annotation for validation. Ultimately, we aim to develop an automatic, cost-effective annotator to create a consistently annotated dataset for training purposes. The next phase of research will focus on instruction tuning an open-access model using proactive snippets from the training dataset, with the goal of improving the model's competence in displaying proactive behaviour.

4. D-Pro Corpus: Proactive Behaviours in Human Dialogues

To understand how proactive behaviours naturally occur in conversation, we examine selected human-human dialogue corpora, focusing on one collaborative phenomenon—proactivity—to investigate its distributional and qualitative features. The study involves corpus annotation and analysis for computational purposes, using empirical data from five task-oriented dialogue corpora: the NESPOLE! Corpus [51], the Ubuntu Chat Corpus [52], the MultiWOZ 2.2 Corpus [53], the JILDA Corpus [54], and the Italian Whatsapp Corpus [5], which provide dialogues from a completely natural setting. From the Whatsapp Corpus, we select excerpts of two-party and multi-party goal-oriented chats. Most dialogues are in Italian, except for those from MultiWOZ 2.2 and excerpts of the Whatsapp Corpus. The resulting corpus, which comprises 151 dialogues, 2,855 turns, and over 6,000 utterances, is named D-Pro Corpus.

We develop a proactivity-oriented annotation schema to label the presence of proactive utterances [55] in conversational turns and to classify proactive behaviours based on the dialogue act's communicative function. The annotation task involves both Agent and Client utterances. The goal is to label an utterance as proactive in its entirety: even if the utterance contains some non-proactive elements, it should still be classified as proactive if it includes any proactive content. Proactive utterances are then further classified according to the dialogue act they convey. To simplify the manual annotation process, we use a restricted set of high-level dialogue acts, chosen from the ISO standard taxonomy by [10], which was developed for annotating dialogue with semantic information. Our objective is to apply the same dialogue act annotation schema across all five sub-corpora, thus requiring high-level dialogue act tags. We select the following dialogue acts, which represent general-purpose communicative functions in [10]'s taxonomy: *inform, suggest, offer, request, instruct.* We further annotate goal-failure situations namely, where one participant cannot fulfil the requests the other has made: we perform this annotation on the intuition that a collaborative participant would show proactive behaviour in order to find a solution to the failure [36]. We also want to investigate the relation between proactive turns and their preceding turns, in order to assess whether all context that is required to motivate proactivity can be found in the adjacent turn or, on the other hand, proactivity has longer dependencies in the dialogue context.

The annotation process is structured into four phases: (i) guidelines creation; (ii) pilot annotation and guidelines revision; (iii) Inter-Annotator Agreement assessment on 15% of the corpus with Cohen's Kappa coefficient [56] between two experienced annotators, resulting in average values of 0.77 for proactive utterance annotation and 0.84 on dialogue act annotation; (iv) extensive D-Pro annotation.

The distributional analysis reveals that, on average, 20% of dialogue turns exhibit proactive behaviour, with the highest rate in the Italian Whatsapp Corpus (36%) and the lowest in MultiWOZ 2.2 (11%). A positive correlation is observed between average turn length and both the percentage of proactive turns and the number of proactive utterances per turn. Regarding dialogue act functions, *inform* is the most prevalent tag (63% of proactive utterances), particularly in the Jilda dataset (73.96%). *suggest* tags are more common in spontaneous dialogues, notably in the Ubuntu Chat Corpus, while *offer* tags are rare. *Request* tags are frequent in the MultiWOZ 2.2 corpus, reflecting its simulated dialogue nature, and *instruct* tags appear often in the Ubuntu Chat Corpus, where Agents guide Clients through troubleshooting. We also explore correlations between linguistic structures and dialogue act annotations, identifying lexical-syntactical patterns linked to proactive behaviour. Qualitative analysis reveals that causal clauses, interrogative clauses, modal verbs, and other frequent structures serve as markers of proactivity: these linguistic markers help distinguish proactive utterances in dialogue, as illustrated by examples from the corpus (see Appendix A).

Additionally, we investigate how proactive utterances are positioned within the flow of a task-oriented dialogue, discussing three aspects: (i) the relation between proactive utterances and goal failures, (ii) the relation between proactivity and the dialogue turn that originates a proactive utterance, and (iii) how proactive utterances are distributed throughout the whole dialogue. Regarding (i), the analysis shows that an average of 57% of failure situations in the D-Pro Corpus prompts proactive utterances, suggesting that proactivity has a significant role in recovering from goal failures. Corpora like NESPOLE! (85% of failures are promptly followed by proactivity) and TO-WhatsApp (71%), which contain higher quantities of proactive behaviour, exhibit fewer failures overall, while MultiWOZ, which shows less proactivity, displays more failures overall. This supports the idea that proactivity helps both in failure recovery and in preventing failures in task-oriented dialogues. For (ii), the analysis of turn adjacency in the D-Pro Corpus reveals that most proactive utterances in MultiWOZ (95.56%) and JILDA (95.09%) follow a reactive utterance within the same turn. In contrast, NESPOLE! and Ubuntu display more non-adjacent proactive utterances, often due to asynchronous messages or backchanneling. This suggests that in certain dialogue contexts, proactive behaviour arises over longer conversational dependencies.

The analysis of (iii) proactivity distribution in dialogues shows that proactive utterances are concentrated in the central portion of task-oriented dialogues, as illustrate in Figure 3. The first and last segments of dialogues have lower proactivity (15% and 10%, respectively), while the central segment has the highest (around 30%). This pattern is consistent across all corpora, except Ubuntu, where proactivity is more evenly distributed probably due to the absence of introductory or closing greetings and the multi-party nature of the interactions.

This study creates a proactivity-centred resource and provides insights on both distributional and qualitative features and on the characteristics of the most suitable corpora to study proactivity and possibly other collaborative strategies.

5. Automatic Proactivity Detection with LLMs

We explore GPT-4o's use as an annotator of proactivity, drawing inspiration from previous research on dialogue coherence annotation and grounding acts. Our study builds on earlier work, particularly focusing on GPT-4o's ability to annotate proactivity, as done in [47]. The model's performance is evaluated using the manually annotated D-Pro Corpus, with human annotations serving as the ground truth. Our research is divided into two phases: (i) whole-dialogue proactivity annotation, and (ii) last-utterance proactivity annotation. In the first phase, we encounter challenges similar to those in dialogue coherence annotation tasks, as classifying each utterance in a full dialogue remains too complex for current models. Additionally, whole-dialogue annotation leads to label imbalance, with only 15% of utterances in the D-Pro Corpus being proactive. To overcome these challenges, we simplify the task by focusing on the final utterance in a dialogue and classifying it as either 'proactive' or 'not_proactive'. To reduce the dialogue context, we extract 4-turn conversational excerpts, as turn-adjacency statistics show that 77.7% of proactive utterances are relevant to the previous turn. We ensure a balanced dataset by collecting an equal number of proactive and non-proactive 4-turn snippets, each ending with a unique utterance. We randomly select 30 snippets for in-context learning, 50 for validation, and 100 for testing from each of the 5 corpora, with snippet selection based on the size of the smallest sub-corpus (MultiWOZ, with 90 proactive utterances).

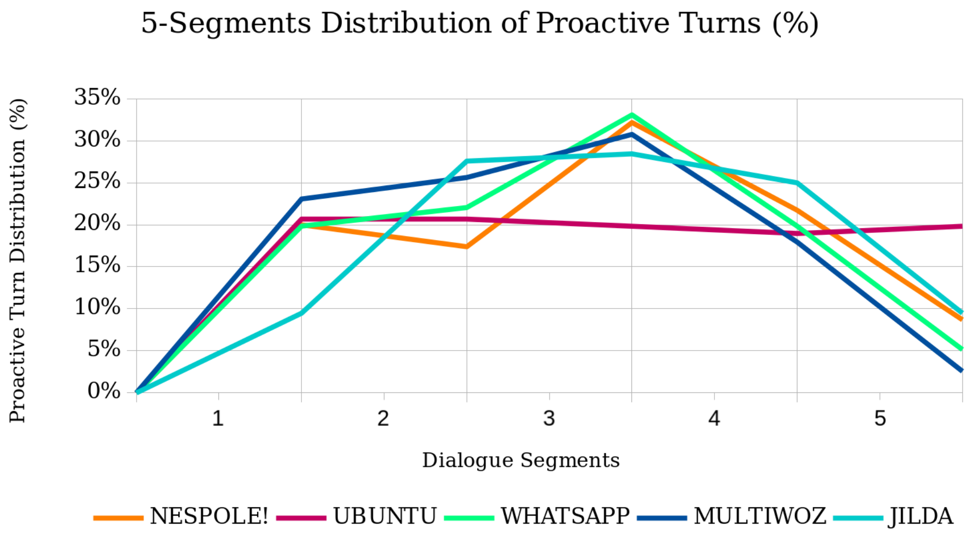
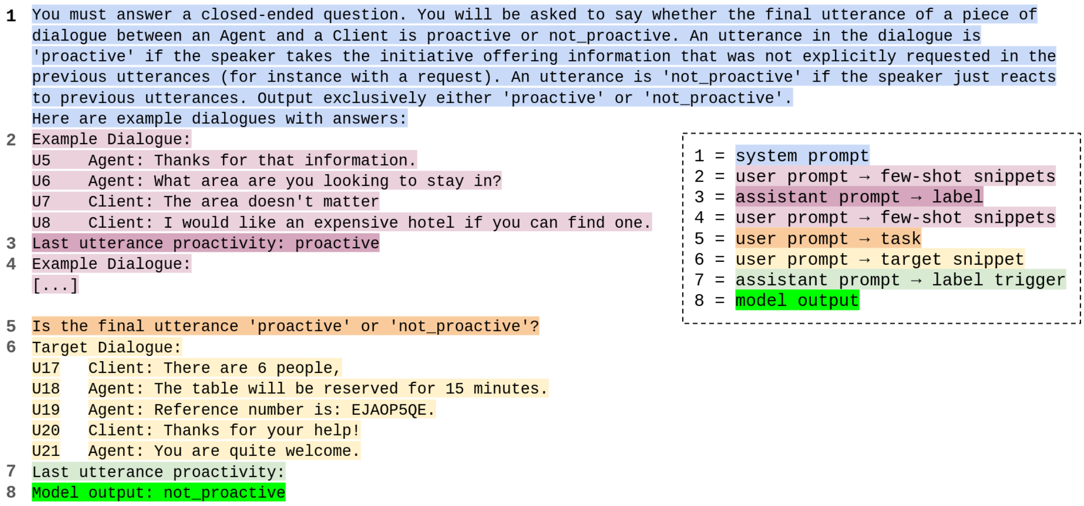


Figure 3: Distribution of proactivity over dialogue turns: each dialogue is divided into five segments and the percentage of proactive turns is computed within each of the five parts, so that dialogues of different lengths are comparable in a coarse-grained analysis.



```
1  You must answer a closed-ended question. You will be asked to say whether the final utterance of a piece of
   dialogue between an Agent and a Client is proactive or not_proactive. An utterance in the dialogue is
   'proactive' if the speaker takes the initiative offering information that was not explicitly requested in the
   previous utterances (for instance with a request). An utterance is 'not_proactive' if the speaker just reacts
   to previous utterances. Output exclusively either 'proactive' or 'not_proactive'.
   Here are example dialogues with answers:
2  Example Dialogue:
   U5    Agent: Thanks for that information.
   U6    Agent: What area are you looking to stay in?
   U7    Client: The area doesn't matter
   U8    Client: I would like an expensive hotel if you can find one.
3  Last utterance proactivity: proactive
4  Example Dialogue:
   [...]

5  Is the final utterance 'proactive' or 'not_proactive'?
6  Target Dialogue:
   U17   Client: There are 6 people,
   U18   Agent: The table will be reserved for 15 minutes.
   U19   Agent: Reference number is: EJAOP5QE.
   U20   Client: Thanks for your help!
   U21   Agent: You are quite welcome.
7  Last utterance proactivity:
8  Model output: not_proactive
```

```
1 = system prompt
2 = user prompt → few-shot snippets
3 = assistant prompt → label
4 = user prompt → few-shot snippets
5 = user prompt → task
6 = user prompt → target snippet
7 = assistant prompt → label trigger
8 = model output
```

Figure 4: Prompt given to the model, divided into system-prompt (blue) and message prompt (violet), target dialogue (yellow) and the output label trigger.

The annotation task is thus transformed into a last-utterance classification task with reduced context, significantly lowering the input prompt length and reducing OpenAI API usage costs. After preparing the training, validation, and test sets, we enter a prompt engineering phase where multiple prompt variations produced by two researchers are tested using the same setting: the same train snippets are used as few-shot examples, the same validation snippets are used to evaluate the model. We experiment variations in the structure of both the system prompt, which contains the general task instructions given to the model, and in the message prompt part, which is further divided into alternating user messages and assistant messages. The latter is the part of the prompt where the model receives few-shot examples (user messages) with the solution to the task (assistant message). The final user/assistant pair contains the target dialogue which is being evaluated by the model at current time. Upon identifying the optimal prompt (Figure 4), we evaluate multiple GPT models. GPT-4o-2024-08-06 emerges as the best performer, offering the best trade-off between cost and performance. We further test the impact of few-shot example order and quantity, with the optimal configuration found to be 12 snippets. First, we test the model using few-shots from individual corpora, yielding varied performance across corpora.

Table 1
Testing the model with few-shot examples taken from all five corpora. Results are given for the best inter-sub-corpus order over five runs.

Metric	Whatsapp	Nespole	Ubuntu	Jilda	MultiWOZ	Average
Accuracy	0.63	0.87	0.66	0.69	0.77	0.72
Precision	0.74	0.88	0.74	0.69	0.78	0.77
Recall	0.4	0.86	0.5	0.6	0.68	0.62
F1 Score	0.52	0.87	0.6	0.69	0.76	0.68
Cohen's Kappa	0.27	0.74	0.32	0.37	0.54	0.44

NESPOLE! achieves the highest accuracy (0.86) and Cohen's Kappa (0.72), followed by MultiWOZ (0.77), while Whatsapp and Ubuntu have the lowest scores, which are expected due to the less structured nature of these corpora. Next, we test the model using few-shot examples from all corpora, aiming to improve performance through transfer learning. The results show that combining corpora as few-shot examples leads to a decrease in performance for Jilda, while other corpora show slight improvements or similar results (Table 1). Finally, we evaluate the model on a cumulative test set of all corpora, using two configurations of few-shot examples: one with 60 snippets and the other with 15. The best results are obtained using 60 snippets in the optimal order, with accuracy reaching 0.71. Overall, our findings suggest that few-shot learning with mixed corpora improves performance in most cases.

We conduct a further experiment to test whether the annotator is context-aware and effectively utilizes the given context to produce its proactivity annotations. Specifically, we investigate the impact of corrupted context on the model's performance by removing or masking the triggering utterance—the key context element prompting the turn that displays proactive behaviour. We hypothesize that corrupting the trigger will increase false positives, as the missing context may lead the model to classify responses as proactive when they are actually reactive. Table 2 shows the results, where accuracy drops from 0.80 in the full context scenario to 0.66 and 0.64 when the triggering utterance is removed or masked, respectively. This drop is mainly due to a rise in false positives (from 2 to 8) and a decrease in true negatives, supporting our hypothesis. Despite the context corruption, the model still outperforms random chance baselines, confirming that solid instruction prompts help maintain performance even with insufficient context. These findings highlight the importance of context integrity for proactivity annotation and demonstrate the model's sensitivity to context corruption, shedding light on its ability to leverage context in decision-making.

Table 2
Proactivity prediction with corrupted dialogue snippets on a sample corpus, MultiWOZ. Highlighted TNs and FPs are statistically different from the test with full context (p-value = 0.04123); results with Trigger Utterance both Empty and Masked are statistically lower (p < 0.01) than in Full Context setting.

	TESTS			BASELINES		
Trigger Utterance	Full Context	Empty	Masked	Full Context	Empty	Masked
True Positives	17	16	15	20	22	23
True Negatives	23	17	17	5	5	5
False Positives	2	8	8	20	20	20
False Negatives	8	9	10	5	3	2
Accuracy	0.80	0.66	0.64	0.50	0.54	0.56
Precision	0.89	0.67	0.65	0.50	0.52	0.53
Recall	0.68	0.64	0.60	0.80	0.88	0.92
F1 Score	0.77	0.65	0.62	0.62	0.66	0.68
Cohen's Kappa	0.59	0.31	0.26	-0.01	0.07	0.11

6. Conclusions and Ongoing Work

Our research towards proactive dialogic agents has resulted so far in: (i) a linguistic and distributional analysis of proactivity in human-human dialogue corpora; (ii) the creation of a manually-curated corpus of annotated proactive dialogues; a resource to (iii) create few-shot examples in an in-context learning approach and (iv) assess an LLM's context-aware capabilities in the annotation of a pragmatic phenomenon; (v) insights on best practices for prompt engineering and parameter setting for our annotation tasks. We aim to explore techniques to further improve the performance of the annotator model, especially in combining dialogue snippets from different corpora. Viable tests include varying the context size and examining whether proactive utterances in the preceding context influence the annotation of the target utterance. Employ the selected model to automatically annotate a large corpus of unlabelled dialogical data, for instance about 100K dialogue snippets.

Next steps along this line of research will focus on selecting the model that offers the best balance between performance, time, and cost efficiency for the task of last utterance proactivity prediction, with a preference for open-source solutions when feasible. To enhance the performance of the annotator model, we plan to explore techniques that allow to more effectively integrate dialogue snippets from different dialogue corpora. We plan on varying the context size to better determine the optimal amount of preceding dialogue needed for accurate annotation. The context size of 4 turns per dialogue snippet was chosen to simplify the annotation task, since the annotation of a whole dialogue yielded poor results. However, selecting a context size somewhere in between 4 turns and the entire dialogue may yield even better results. Given that the GPT-4o annotator model output higher scores on the NESPOLE! dialogues, and that the NESPOLE! dialogues have longer turns than the other four investigated corpora (2.85 utterances per NESPOLE! turn, versus 2.11 utterance per D-Pro turn on average), context length may be a key variable in the model's performance. Other viable strategies include investigating whether the presence of proactive utterances in the preceding context significantly influences the annotation of the target utterance.

Once the optimal model is selected and all parameters are set, we plan to use it to automatically annotate a large corpus of unlabelled dialogical data, targeting approximately 100,000 dialogue snippets. This step involves also the challenge of processing the unlabelled dialogue data in order to automatically segment turns into utterances, a task that may be addressed with the approach outlined in [57].

The final step will involve leveraging the corpus automatically labelled by the annotator model instruction tune an open model, ultimately enhancing its pragmatic competence and paving the way for a more proactive and contextually aware dialogic AI system.

Acknowledgement

Bernardo Magnini was partially supported by the PNRR MUR project PE0000013-FAIR (Spoke 2).

References

[1] G. Leech, 136 Pragmatics and Dialogue, in: The Oxford Handbook of Computational Linguistics, Oxford University Press, 2005. URL: https://doi.org/10.1093/oxfordhb/9780199276349.013.0007. doi:10.1093/oxfordhb/9780199276349.013.0007. arXiv:https://academic.oup.com/book/0/chapter/293281482/chapter-ag-pdf/44512892/book_34563_section_293281482.ag.pdf.

[2] R. Fernández, Dialogue, in: The Oxford Handbook of Computational Linguistics, Oxford University Press, 2022, pp. 179–204. URL: https://doi.org/10.1093/oxfordhb/9780199573691.013.25_update_001. arXiv:https://academic.oup.com/book/0/chapter/358148873/chapter-pdf/45719672/oxfordhb-9780199573691-e-25.pdf.

[3] T. Fong, C. Thorpe, C. Baur, Collaboration, dialogue, human-robot interaction, in: Robotics Research, Springer, 2003, pp. 255–266.

[4] F. Nothdurft, S. Ultes, W. Minker, Finding appropriate interaction strategies for proactive dialogue systems—an open quest, in: Proceedings of the 2nd European and the 5th Nordic Symposium on Multimodal Communication, volume 110, Citeseer, 2014, pp. 73–80.

[5] F. Hewett, Sequential Organisation in WhatsApp Conversations., Tesi di laurea triennale non pubblicata, Libera Università di Berlino, semestre estivo, 2017.

[6] P. Grice, Logic and conversation, in: Speech acts, Brill, 1975, pp. 41–58.

[7] P. Grice, Studies in the Way of Words, Harvard University Press, 1989.

[8] J. L. Austin, How to do things with words, William James Lectures, Oxford university press, 1962. URL: http://scholar.google.de/scholar.bib?q=info:xI2JvixH8_QJ:scholar.google.com/&output=citation&hl=de&as_sdt=0,5&ct=citation&cd=1.

[9] D. R. Traum, E. A. Hinkelman, Conversation acts in task-oriented spoken dialogue, Computational intelligence 8 (1992) 575–599.

[10] H. Bunt, J. Alexandersson, J. Carletta, J.-W. Choe, A. C. Fang, K. Hasida, K. Lee, V. Petukhova, A. Popescu-Belis, L. Romary, et al., Towards an iso standard for dialogue act annotation, in: Seventh conference on International Language Resources and Evaluation (LREC'10), 2010.

[11] H. Bunt, Y. Girard, Designing an open, multidimensional dialogue act taxonomy, Proceedings of the 9th Workshop on the Semantics and Pragmatics of Dialogue (2005).

[12] H. Bunt, Dimensions in dialogue act annotation, in: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), European Language Resources Association (ELRA), Genoa, Italy, 2006. URL: http://www.lrec-conf.org/proceedings/lrec2006/pdf/428_pdf.pdf.

[13] P. Brown, S. C. Levinson, S. C. Levinson, Politeness: Some universals in language usage, volume 4, Cambridge university press, 1987.

[14] E. Goffman, Interaction ritual (1967), 1972.

[15] F. Bargiela-Chiappini, Face and politeness: New (insights) for old (concepts), Journal of pragmatics 35 (2003) 1453–1469.

[16] H. Giles, D. M. Taylor, R. Bourhis, Towards a theory of interpersonal accommodation through language: some canadian data1, Language in society 2 (1973) 177–192.

[17] S. M. Burt, Code choice in intercultural conversation: Speech accommodation theory and pragmatics, Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA) 4 (1994) 535–559.

[18] C. M. Scotton, Codeswitching as indexical of social negotiations, Codeswitching: Anthropological and sociolinguistic perspectives (1988) 151–186.

[19] C. Gallois, T. Ogay, H. Giles, Communication accommodation theory: A look back and a look ahead, in: Theorizing about intercultural communication, Thousand Oaks: Sage, 2005, pp. 121–148.

[20] H. H. Clark, E. F. Schaefer, Collaborating on contributions to conversations, Language and cognitive processes 2 (1987) 19–41.

[21] H. H. Clark, E. F. Schaefer, Contributing to discourse, Cognitive science 13 (1989) 259–294.

[22] H. Clark, Grounding in communication, Perspectives on socially shared cognition/American Psychological Association (1991).

[23] H. H. Clark, Using language, Cambridge university press, 1996.

[24] D. R. Traum, A computational theory of grounding in natural language conversation., Technical Report, Rochester Univ NY Dept of Computer Science, 1994.

[25] T. Visser, D. Traum, D. DeVault, R. op den Akker, A model for incremental grounding in spoken dialogue systems, Journal on Multimodal User Interfaces 8 (2014) 61–73.

[26] M. Purver, J. Ginzburg, P. Healey, On the means for clarification in dialogue, in: Current and new directions in discourse and dialogue, Springer, 2001, pp. 235–255. URL: https://www.aclweb.org/anthology/W01-1616/.

[27] M. Purver, P. Healey, J. King, J. Ginzburg, G. J. Mills, Answering clarification questions, in: Proceedings of the Fourth SIGdial Workshop of Discourse and Dialogue, 2003, pp. 23–33.

[28] M. Purver, Clarie: Handling clarification requests in a dialogue system, Research on Language and Computation 4 (2006) 259–288.

[29] S. Stoyanchev, A. Liu, J. Hirschberg, Towards natural clarification questions in dialogue systems, in: AISB symposium on questions, discourse and dialogue, volume 20, 2014.

[30] M. Aliannejadi, J. Kiseleva, A. Chuklin, J. Dalton, M. Burtsev, Convai3: Generating clarifying questions for open-domain dialogue systems (clariq), arXiv preprint arXiv:2009.11352 (2020).

[31] L. Shelley, F. Gonzalez, Back channeling: Function of back channeling and l1 effects on back channeling in l2, Linguistic Portfolios 2 (2013) 9.

[32] M. F. Jung, J. J. Lee, N. DePalma, S. O. Adalgeirsson, P. J. Hinds, C. Breazeal, Engaging robots: easing complex human-robot teamwork using backchanneling, in: Proceedings of the 2013 conference on Computer supported cooperative work, 2013, pp. 1555–1566.

[33] M. Murray, N. Walker, A. Nanavati, P. Alves-Oliveira, N. Filippov, A. Sauppe, B. Mutlu, M. Cakmak, Learning backchanneling behaviors for a social robot via data augmentation from human-human conversations, in: Conference on Robot Learning, PMLR, 2022, pp. 513–525.

[34] A. Fetzer, Reformulation and common grounds, in: Lexical markers of common grounds, Brill, 2006, pp. 159–181.

[35] H. Giles, T. Ogay, et al., Communication accommodation theory, Explaining communication: Contemporary theories and exemplars (2007) 293–310.

[36] V. Balaraman, B. Magnini, Pro-active systems and influenceable users: Simulating pro-activity in task-oriented dialogues, Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue (2020).

[37] L. Liao, G. H. Yang, C. Shah, Proactive conversational agents (2023).

[38] Y. Deng, W. Lei, W. Lam, T.-S. Chua, A survey on proactive dialogue systems: Problems, methods, and prospects, arXiv preprint arXiv:2305.02750 (2023).

[39] Y. Deng, L. Liao, L. Chen, H. Wang, W. Lei, T.-S. Chua, Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration, arXiv preprint arXiv:2305.13626 (2023).

[40] M. Kraus, N. Wagner, W. Minker, Effects of proactive dialogue strategies on human-computer trust, in: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, 2020, pp. 107–116.

[41] P.-M. Strauss, W. Minker, Proactive spoken dialogue interaction in multi-party environments, Springer, 2010.

[42] M. Kraus, F. Fischbach, P. Jansen, W. Minker, A comparison of explicit and implicit proactive dialogue strategies for conversational recommendation, in: Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 429–435.

[43] M. L'Abbate, U. Thiel, T. Kamps, Can proactive behavior turn chatterbots into conversational agents?, in: IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IEEE, 2005, pp. 173–179.

[44] M. Kraus, N. Wagner, W. Minker, Prodial–an annotated proactive dialogue act corpus for conversational assistants using crowdsourcing, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 3164–3173.

[45] M. Kraus, N. Wagner, N. Untereiner, W. Minker, Including social expectations for trustworthy proactive human-robot dialogue, in: Proceedings of the 30th ACM conference on user modeling, adaptation and personalization, 2022, pp. 23–33.

[46] M. Kraus, N. Wagner, R. Riekenbrauck, W. Minker, Improving proactive dialog agents using socially-aware reinforcement learning, in: Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization, 2023, pp. 146–155.

[47] O. Shaikh, K. Gligorić, A. Khetan, M. Gerstgrasser, D. Yang, D. Jurafsky, Grounding gaps in language model generations, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2024, pp. 6279–6296.

[48] T. Labruna, S. Brenna, A. Zaninello, B. Magnini, Unraveling chatgpt: A critical analysis of ai-generated goal-oriented dialogues and annotations, arXiv preprint arXiv:2305.14556 (2023).

[49] F. Huang, H. Kwak, J. An, Is chatgpt better than human annotators? potential and limitations of

chatgpt in explaining implicit hate speech, ArXiv abs/2302.07736 (2023).

[50] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).

[51] N. Mana, S. Burger, R. Cattoni, L. Besacier, V. MacLaren, J. McDonough, F. Metze, The nespole! voip multilingual corpora in tourism and medical domains, in: Eighth European Conference on Speech Communication and Technology, 2003.

[52] D. C. Uthus, D. W. Aha, The ubuntu chat corpus for multiparticipant chat analysis, in: 2013 AAAI Spring Symposium Series, 2013.

[53] X. Zang, A. Rastogi, S. Sunkara, R. Gupta, J. Zhang, J. Chen, MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines, in: Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, Association for Computational Linguistics, Online, 2020, pp. 109–117. URL: https://www.aclweb.org/anthology/2020.nlp4convai-1.13. doi:10.18653/v1/2020.nlp4convai-1.13.

[54] I. Sucameli, A. Lenci, B. Magnini, M. Simi, M. Speranza, Becoming jilda, in: J. Monti, F. Dell'Orletta, F. Tamburini (Eds.), Proceedings of the Seventh Italian Conference on Computational Linguistics CLIC-it 2020, volume 2769 of *CEUR Workshop Proceedings*, CEUR-WS, Bologna, 2020. URL: http://ceur-ws.org/Vol-2769/paper_69.pdf.

[55] D. Traum, Issues in multiparty dialogues, in: Workshop on Agent Communication Languages, Springer, 2003, pp. 201–211.

[56] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, biometrics (1977).

[57] M. Frohmann, I. Sterner, I. Vulić, B. Minixhofer, M. Schedl, Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation, arXiv preprint arXiv:2406.16678 (2024).

[58] R. Lowe, N. Pow, I. V. Serban, J. Pineau, The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems, in: Proceedings of the SIGDIAL 2015 Conference, 2015, pp. 285–294.

A. Patterns frequently co-occurring with proactivity in D-Pro: example dialogue excerpts.

Example: causal clause

a: **U6** Il mio sogno sarebbe quello di fare l'insegnante
U7 [PRO][INFORM] perché mi piace lavorare con i bambini e ragazzi.[1]

Example: interrogative clause

a: **U8** I would like an expensive hotel if you can find one.
b: **U9** The express by holiday inn cambridge is located in the east and meet your criteria.
U10[PRO][OFFER] Shall I book you a room?[2]

Example: modals

b: **U13** I've found several restaurants that are located in the Centre with a moderate price range.
U14 [PRO][OFFER] May I recommend a British restaurant called the Oak Bistro?[3]

Example: connectives

b: **U41** trus, ovvero: hai installato ubuntu "dentro" windows?

[1] Example taken from the JILDA Corpus [54].

[2] Example taken from the MultiWOZ 2.2 Corpus [53].

[3] Example taken from the MultiWOZ 2.2 Corpus [53].

a: **U42** mi sa che hai ragione...anzi si...
U43 **[PRO][INFORM]** però mi sembra di ricordare che il disco in qualche modo me lo ha fatto partizionare lo stesso...[4]

Example: pattern ”I suggest/recommend that you... / ti consiglio di...”

a: **U30** Mi potresti fornire informazioni sull'altra proposta di lavoro?
b: **U31** certo, attendi solo un momento per favore
U32 **[PRO][SUGGEST]** ti consiglio di informarti comunque presso la Munus s.r.l.[5]

Example: pattern ”try doing... / prova a...”

a: **U13** marcotux, puoi spiegarmi come si fa?
b: **U14** Under_Flea, provo a vedere se esiste in pacchetto
U15 **[PRO][SUGGEST]** prova a vedere nel gestore pacchetti se c'è lastfm.[6]

[4] Example taken from the Ubuntu Chat Corpus [58].

[5] Example taken from the JILDA Corpus [54].

[6] Example taken from the JILDA Corpus [54].