INNOVATION IN LANGUAGE LEARNING International Conference

Designing Task-based Chatbots for L2 Pragmatics Learning: Experiments and Preliminary Studies

Federica Turco[1, 2], Giuliana Franceschinis[2], María Joaquina Valero Gisbert[3]

[1]University of Eastern Piedmont, Italy
[2]University of Modena and Reggio Emilia, Italy
[3]University of Parma, Italy

Abstract

In Italian higher education, pragmatics teaching remains underdeveloped and is often addressed through traditional, form-focused approaches [1]. The need to expand language practice and pragmatic awareness, while increasing university students' engagement, reducing in-class time, and providing learners with personal tutors, has led to the development of LLM-based chatbots, which are currently among the most promising tools in Technology-mediated Second Language [2] learning. The Task-Based Language Teaching [3] approach, within the CALL [4] framework, can be leveraged to design chatbots able to enhance the development of pragmatic competence in Italian L2 task-based scenarios. Chatbots can structure task-based service-encounter interactions and act as conversational partners that provide immediate, context-sensitive feedback aligned with the Common European Framework of Reference for Languages (CEFR) descriptors [5]. Preliminary experiments analysed in this contribution compare progressively more structured LLM-based conversational systems, moving from prompt-based ChatGPT interactions to Custom GPTs configured through increasingly specific instructions and external knowledge. The latest configurations, developed under the name "PragmaBot", use configuration files to structure conversational behaviour, situational knowledge, and pedagogical feedback. The results show that more specific prompting and structured knowledge improve several aspects of the interaction, including role consistency, register, task progression and feedback, but do not ensure stable and predictable pragmatic behaviour. Pragmatics is highly context-dependent and socially situated, which makes it particularly difficult to shape in black-box, general-purpose chatbots, typically more effective at handling linguistic form than pragmatic use. These findings motivate the subsequent development of a more controllable agentic application under design, PragmAI, guided by an explicit representation of pedagogical-pragmatic knowledge and designed to separate and coordinate the different functions involved in task-based conversational interaction. The agent would also refer to a database of learners' characteristics, needs and difficulties to foster a personalised interaction for university students in the aforementioned communicative context.

Keywords: *Pragmatics; Chatbot; TBLL; CALL; Technology-mediated Second Language Pragmatics*

1. Introduction

Artificial Intelligence is increasingly reshaping language teaching and learning, supporting the development of Artificial Intelligence-Assisted Language Learning (AIALL), a recent area within the Computer-Assisted Language Learning (CALL) framework that investigates how AI tools and techniques can support language-learning processes [6]. Among these emerging technologies, Large Language Model (LLM)-based conversational agents offer particular potential to extend opportunities for interaction and oral practice beyond teacher-led classroom activities. Within this domain, the present research focuses on Technology-Mediated Second Language Pragmatics [2], specifically on teaching and learning Italian L2 pragmatics for international university students through task-based conversational interaction. Pragmatic competence concerns the ability to use and interpret language appropriately in context and therefore entails not only linguistic knowledge, but also sensitivity to interlocutor roles, communicative intentions, social relationships and sociocultural norms. It should also be related explicitly to the linguistic resources through which spoken interaction is constructed: pragmatic phraseological units, conversational routines, discourse markers, forms of address, mitigation and intensification strategies, as well as features of colloquial speech, including colloquial lexical choices, and of oral interaction (e.g., discourse markers, formulaic sequences, ellipsis, reformulation, hesitation, turn-management devices, and prosodic cues [7]). For L2 Italian learners, knowing the grammatical form of an utterance does not necessarily entail knowing when, with whom, or for what

interactional purpose it is appropriate [8]. Its development is therefore particularly relevant to successful social interaction, although pragmatic instruction remains comparatively underrepresented in language curricula and is often still approached through predominantly form-focused practices, including in Italian higher education [1]. Moreover, pragmatics presents a specific challenge for AI-supported learning because appropriate language use is highly context-dependent and socially situated: the effectiveness of an utterance cannot be evaluated solely in terms of grammatical or semantic correctness, but depends on the communicative situation in which it occurs. For this reason, the research adopts a Task-Based Language Teaching/Learning (TBLT/TBLL) perspective [3; 9], in which language is used to accomplish meaningful and realistic communicative goals. This approach appears particularly suitable for pragmatic learning and can be combined with digital technologies through Technology-mediated TBLT, an area that is still relatively recent and under-researched [10].

The contribution reports a preliminary stage of a doctoral research project investigating how an oral LLM-based conversational system can support the development of pragmatic competence in Italian L2 among international university students. Rather than assessing learners' outcomes at this stage, the study centres on the chatbot's design and behaviour. It investigates whether currently available customisable LLM-based systems can be effectively guided through increasingly structured prompting and externally provided knowledge to sustain task-based interactions that are pedagogically coherent and pragmatically appropriate. Preliminary experimentation therefore identifies recurrent problems in the chatbot's management of interaction and pragmatic behaviour, determining which requirements are hard to handle through prompting alone. These limitations motivate exploring a more controllable solution, namely an agentic application, in which an explicit representation of pragmatic knowledge and more structured interactional procedures can support conversational behaviour [11]. The experiments discussed in this paper consequently represent an intermediate step between the initial customisation of a general-purpose LLM and the subsequent development of PragmAI, the agentic system envisaged within the wider research project.

The contribution examines the conditions required for meaningful pedagogical integration of GenAI to support pragmatic competence. By connecting AIALL, L2 pragmatics and TBLT, the study researches where general-purpose conversational technologies run into difficulties when applied to a highly situated learning objective, and how the problems observed during testing can inform the design of more dedicated educational systems. In this respect, the work addresses a broader question central to technology-enhanced language education: how an LLM could be controlled to support a specific language-learning objective in conversation with a learner.

2. Methodology

The preliminary experimentation followed an iterative design-and-test process to identify which type of LLM-based conversational system could more reliably support task-based practice of Italian L2 pragmatics. Three progressively more controlled implementation options were considered: (i) a standard ChatGPT conversation guided through system prompting, (ii) customised GPTs combining prompting with externally provided knowledge, and (iii) an agentic application. This study reports experiments with the first two options from ChatGPT's paid version; the third, currently under development on Gemini, represents the subsequent design direction emerging from their limitations. The experiments focused primarily on restaurant service encounters, since this domain provides situated communicative tasks with identifiable roles, phases and pragmatic conventions, while allowing variation in degree of formality and service organisation. Across the experiments, the researcher acted as the simulated learner, adopting different proficiency and behavioural profiles according to the purpose of each test. Moreover, chatbot behaviour was assessed qualitatively against the expected interactional and pragmatic features defined in the prompts and configuration files, including role consistency, register, forms of address, task progression, conversational naturalness, and feedback behaviour.

The first exploratory condition used a standard ChatGPT conversation initiated with a general prompt defining the role-play in tests focused on increasingly situated restaurant interactions. In the restaurant condition, ChatGPT was instructed to assume the role of the server while the researcher played the customer, without supplying an external knowledge base. This preliminary phase aimed at observing the general-purpose model's default conversational tendencies before introducing stronger constraints. It revealed the need for greater control over contextual interpretation, forms of address, turn progression, and the concrete advancement of the simulated interaction.

The second condition therefore involved a series of Custom GPTs, configured through increasingly detailed instructions and uploaded knowledge. Early versions combined prompts specifying context,

background, objectives, and conversational behaviour with materials related to TBLT/TBLL, second-language teaching and the CEFR proficiency levels. Subsequent configurations refined the role description and pedagogical objectives and added new functionalities, along with a document on the design and requirements analysis for a conversational TBLL chatbot that explicitly defined the expected behaviour of the conversational tutor, including role coherence, task progression, pragmatic support, feedback, and interactional constraints. However, this progressive enrichment also generated a methodological problem: information on role, task structure, correction, and feedback was distributed across the general prompt, theoretical materials, and design documents. This redundancy sometimes created overlapping or partially contradictory instructions, making it difficult to determine which source should govern the chatbot's behaviour; for example, it could alternate forms of address within the same conversation or handle different stages of restaurant interaction differently. This observation motivated a shift from simply increasing the quantity of information provided to the model to more explicitly structuring and hierarchizing knowledge.

For this purpose, the next Custom GPT configurations were specialised through configuration files. Markdown [12] was selected as a semi-structured representation format, making the behavioural and pedagogical knowledge readable, modular and directly editable without specific tools. In the first configuration, developed for an elegant restaurant scenario, knowledge was divided into three complementary files: the first one defines the conversational role and its behaviour; the second one defines the restaurant scenario, interactional phases, and situational constraints; finally, the third one defines the post-task pedagogical evaluation and feedback. The files were initially tested in PragmaBot alongside the design-and-requirements document. Because different experts may produce these materials and express them relatively freely in natural language, overlaps or conflicting instructions may emerge; this issue also appeared in the present experimentation. PragmaBot 1 therefore reorganised the general coordinating prompt, removed the additional design document from the operational knowledge base, and made the three Markdown files the chatbot's primary reference to specialise the system.

A further step addressed the limited transferability of scenario-specific rules. The three-file architecture for the elegant restaurant encoded features (such as mandatory use of the courtesy formula constituted by the third-person pronoun "Lei", a specific service sequence and a particular distribution of roles) that could not be generalised to a pizzeria, a self-service restaurant or a university cafeteria. Therefore, an abstract Markdown configuration was created to separate invariant conversational and pedagogical rules from scenario-dependent information. The resulting file contains only general rules concerning role maintenance, task progression, student initiative, pragmatic error management, separation between role-play and feedback, and evaluation criteria; elements such as register, form of address, concrete roles, menu, payment procedure, and task sequence are explicitly delegated to future context-specific files. The single abstract configuration was thus loaded into a new Custom GPT, PragmaBot-Astratto, and tested across four restaurant-related scenarios: elegant restaurant, pizzeria, self-service, and university cafeteria. In the experiments reported in this paper, the abstract schema was intentionally tested without dedicated context files to investigate how much the model's implicit knowledge of restaurant-related situations could complement abstract behavioural rules, and to identify which aspects instead required explicit situational knowledge. This condition also provides the basis for a subsequent controlled comparison with configurations including dedicated context files, to examine whether more explicit contextual information increases the appropriateness, variability, and controllability of the interaction. Such contextual knowledge should preferably be grounded in appropriately annotated corpora or other empirically documented authentic interactions [13], particularly when pragmatic features become explicit learning targets, so that the system relies on attested patterns of interaction rather than on the LLM's statistically plausible representations.

3. Results: Prompt-and-Markdown based Configurations

3.1 Prompt Based Configurations

Some preliminary oral interaction tests were affected by speech-recognition and transcription errors, making it difficult to distinguish problems in conversational and pragmatic behaviour from those introduced by the voice interface. The present phase therefore focuses primarily on the system's conversational behaviour, using written interaction where appropriate. This choice temporarily excludes features specific to spoken interaction, including timing, hesitation, overlapping speech, backchanneling,

and pronunciation-related issues, which will require separate evaluation once the conversational configuration has been sufficiently stabilised.
An initial restaurant role-play used the paid version of ChatGPT with a general prompt defining server and customer roles. Although the model maintained the setting, it autonomously inferred unspecified pragmatic variables, such as interpersonal familiarity and the informal “tu”, showing that role assignment alone was insufficient to control context-dependent behaviour.
Three subsequent Custom GPT configurations progressively increased specification during the second experimental phase. PragmAI, running automatically on GPT-5.2, combined an extensive prompt with theoretical materials on TBLT, second-language teaching, and the CEFR [5]. PragmAI 2, running on the more advanced GPT-5.6 Sol, retained this knowledge base but added more operational instructions on context, pedagogical objectives, role, and conversational behaviour. This improved role entry, use of formal “Lei”, interpretation of indirect moves, and turn progression, although verbosity, excessive deference, incomplete action execution, and unreliable feedback activation persisted. Pragmabot, also running on GPT-5.6 Sol, retained the previous prompt while introducing technical changes (data analysis enabled and model-improvement data sharing disabled) and adding a design-and-requirements document that specified operational aspects such as role consistency, task progression, pragmatic support, interactional brevity, and feedback. However, theoretical and methodological sources did not consistently translate into the required conversational behaviour, and heterogeneous materials sometimes made prioritisation more difficult. This led to a shift from enlarging the prompt and knowledge base to extracting operational principles into dedicated configuration files.

3.2 Markdown Based Configurations

These limitations were addressed by restructuring the knowledge supplied to the Custom GPT through configuration files, using the same paid ChatGPT environment and technical settings: as in the previous tests, since the configuration specified no recommended model, the platform automatically ran GPT-5.6 Sol.
The first configuration was developed for an elegant restaurant and divided the relevant information into three files (conversational role, situational context, and pedagogical evaluation), and also included the design and requirements document. The general prompt largely retained the previous configuration, with a minor revision allowing linguistic and pragmatic help when necessary.
Four oral interactions were selected as particularly relevant for analysing this configuration. The simulated learner profiles varied in proficiency and behaviour, and included a B1 learner with stronger comprehension than oral production, advanced learners approximately within the C1-C2 range, and interactions in which pragmatic errors were deliberately introduced in order to observe both the chatbot’s conversational reaction and its subsequent evaluation: the first B1 simulation, for example, was explicitly designed around good comprehension but weaker oral production. Initial tests with PragmaBot customised GPT showed that this explicit decomposition improved access to scenario-specific information, but also exposed inconsistencies between the general GPT prompt and the uploaded files. For example, instructions concerning the conversational role and the subsequent Tutor function were distributed across different configuration components, without a sufficiently explicit hierarchy specifying when each should apply; similar overlaps concerned task progression, correction, and feedback: the issue therefore concerned the organisation and prioritisation of instructions rather than their absence.
Tests of this first Markdown-based configuration nevertheless showed progressive improvements in some behaviours: the system increasingly recovered the correct menu information, maintained “Lei” more consistently and adhered more closely to the intended restaurant scenario. However, six recurrent difficulties remained: misunderstanding or overinterpreting unclear input, losing track of the current task phase, excessive politeness and unnecessary conversational expansion, inconsistent representation of physical actions, unreliable or overly positive final feedback, and limited naturalness in turn-taking. The latter was particularly evident in the absence of genuine backchannel behaviour: expressions such as “certo” (“certainly”), “capisco” (“I understand”) or “perfetto” (“perfect”) appeared as the beginning of complete chatbot turns rather than as short listener signals overlapping or supporting the speaker’s ongoing turn. The interaction therefore remained dialogic but not fully comparable to spontaneous conversation.
These findings led to PragmaBot 1, which shortened the general prompt, prioritised the three configuration files, and removed the additional theoretical and design materials. The revised instructions emphasised brief, natural turns; professional but not excessively deferential behaviour; cautious

interpretation of unclear input; avoidance of explicit correction during role-play; and a clear transition to Tutor feedback after task completion.

Two written tests were then conducted with PragmaBot 1 in order to temporarily remove speech recognition as an intervening variable, simulating a B2 and a C1 learner, respectively. The results were markedly more stable: the chatbot retrieved the menu correctly, maintained the expected register, produced shorter, less servile turns, followed the task phases more consistently, and generated feedback that more closely followed the structure specified in the file describing the expected type of feedback. The feedback also became more closely anchored to actual learner productions rather than generic praise.

Some problems nevertheless persisted. The chatbot could still infer information that had not been provided, such as assigning a masculine gender to the customer from a surname alone. Action representation improved but was not fully systematic, and the distinction between the form of address used by the server and that adopted by the Tutor still required explicit clarification. Moreover, because these tests were conducted in writing, the improvement cannot be attributed exclusively to the revised configuration: removing automatic speech recognition also eliminated a major source of ambiguity. The written tests therefore demonstrate that a clearer hierarchy between prompt and configuration files can increase consistency when the linguistic input itself is stable, but they do not yet establish equivalent reliability in spoken interaction.

The next step was to abstract the general behavioural rules from the elegant-restaurant scenario. PragmaBot-Astratto was also configured in the paid ChatGPT environment using GPT-5.6 Sol, but both its instructions and its knowledge base were simplified. Instead of referring to three scenario-specific files, the prompt mainly served an operational framing function, recalling the need for brief, natural turns; appropriate conversational progression; learner initiative; cautious management of unclear input; and a clear separation between role-play and feedback. A single configuration file was uploaded because it was designed to retain only invariant principles concerning role maintenance, conversational progression, learner initiative, management of pragmatic problems and post-task feedback. Scenario-dependent information (such as formality, forms of address, menu, roles, service organisation, payment procedure, and specific task phases) was intentionally excluded.

The resulting PragmaBot-Astratto custom GPT was tested by simulating different learner profiles acting in four written scenarios specified at the beginning of each interaction, without providing any dedicated context file or additional scenario-specific information: an autonomous C2 learner in the elegant restaurant, also characterised by a specific dietary requirement; a C1 learner in the pizzeria, whose interaction was deliberately characterised by a relatively formal register; a highly autonomous B1 learner in the self-service scenario; and a university cafeteria interaction. The abstract schema was deliberately tested without dedicated context files in order to investigate how far the model's implicit knowledge of restaurant-related situations could complement general behavioural rules, and which elements instead required explicit situational knowledge. This condition also provides a basis for a subsequent comparison with configurations that include dedicated context files specifying scenario-dependent features, such as register and forms of address, service organisation, task sequence, menu information, and payment procedures.

A central result emerged consistently across the four tests: the chatbot correctly recognised that the abstract schema lacked the situational information needed for a complete role-play and generally avoided inventing it, as the instructions required. In the elegant restaurant, for example, the absence of a menu led the system to repeatedly state that it needed to check with the kitchen rather than generating unsupported dishes. In the self-service and university cafeteria, it explicitly acknowledged the absence of the relevant context file: in the latter, it refrained from inventing both menu and prices, depending in both cases on the student to supply missing elements. This behaviour suggests that the abstract schema can serve as a reusable regulatory layer. However, because LLMs may already encode knowledge of familiar service encounters, these no-context tests are not sufficient to determine which situational information must be provided explicitly. A direct comparison between the abstract schema alone and the same schema combined with dedicated context files is therefore needed.

This context dependency was also visible in conversational initiative: in the pizzeria the learner introduced almost all task phases, while in the cafeteria the learner had to supply missing menu and price information. The role-play could therefore be completed with highly autonomous simulated learners, but the same configuration would provide less scaffolding to lower-proficiency learners. Previous tests with PragmaBot 1 suggest that explicit, scenario-specific information can support more controlled interaction, for example in menu retrieval and task progression. However, since the prompt and knowledge configuration also changed across versions, these results cannot isolate the effect of

contextual knowledge alone. A controlled comparison should therefore keep the prompt and abstract schema constant while varying only the presence of a dedicated context file.
The experiments therefore produced two complementary findings: first, explicitly structured and hierarchized knowledge substantially improves a Custom GPT's controllability, particularly with respect to register, task progression, verbosity, and feedback; second, general behavioural rules cannot replace scenario-specific knowledge. The most promising configuration emerging from the tests is consequently not the abstract schema alone, but a two-level structure in which a stable regulatory layer is combined with separate context files containing the concrete information required by each communicative situation. Even under this improved representation, however, the Custom GPT still assigns multiple functions to the same model (input interpretation, conversational-state management, pragmatic decision-making, task advancement, and final evaluation), leaving their activation insufficiently predictable. This residual limitation provides the main empirical motivation for an agentic architecture, as discussed in the following section.

4. Discussion

The experiments indicate that, with more detailed instructions, the main challenge in designing an LLM-based chatbot for L2 pragmatics is ensuring that relevant knowledge is consistently and contextually applied in the appropriate moment of the interaction. Across configurations, increasingly specific prompting and additional knowledge generally improved role maintenance, register, task progression, and feedback. However, these improvements were not linear or stable: a modification that enhanced one dimension could weaken another, and behaviours correctly produced in one interaction could disappear in the next. This instability is particularly relevant for pragmatics, where appropriateness depends not only on linguistic form, but also on interlocutor roles, the participants' roles and relationship, the expected register and form of address, and the current stage of the interaction.
The results therefore support the initial assumption that pragmatic competence is more difficult to operationalise than morphosyntactic or lexical correctness within a general-purpose LLM. The system often produced grammatically correct and semantically coherent utterances that nevertheless appeared pragmatically implausible in the simulated context. Examples of this mismatch include excessive politeness, unnecessary comments, premature questions, inappropriate forms of address and insufficiently natural turn progression. The problem was especially evident when the model interpreted politeness quantitatively, by accumulating deferential formulas, rather than contextually. In an elegant restaurant, for instance, professional behaviour requires discretion and appropriate interpersonal distance rather than constant reassurance or availability. These observations confirm the need to evaluate chatbot performance in terms of pragmatic appropriateness and interactional effectiveness, rather than linguistic correctness alone. Although the expected register and behaviour were specified, the model did not always maintain them throughout longer interactions. This suggests that relevant contextual constraints may need to be reactivated during the conversation, a function that an agentic architecture could support.
The experiments also highlight the importance of knowledge representation. Moving from overlapping prompts and sources to a clearer hierarchy of structured configuration files improved the system's organisation and stability, suggesting that the scope, priority, and internal consistency of knowledge matter as much as its quantity. The abstract Markdown schema further distinguished general pragmatic-interactional rules from scenario-specific information, such as register, forms of address, service sequence, and interlocutor roles. However, the tests also showed that general behavioural rules cannot replace contextual knowledge: without it, the chatbot either became overly cautious or relied excessively on learner initiative. This is particularly relevant pedagogically, since lower-proficiency learners may be less able to compensate for missing system knowledge. A suitable configuration should therefore combine a stable regulatory layer with explicit scenario-specific knowledge, supporting learner autonomy without transferring responsibility for task progression to the learner.
A further issue concerns combining conversational and pedagogical functions within the same Custom GPT. The tests repeatedly showed difficulties transitioning from role-play to feedback, timing feedback activation and, in some cases, attributing errors correctly. Feedback quality also improved with the more structured configurations, becoming more closely aligned with the expected format and with the learner's actual productions; however, its reliability remained limited by the difficulty of separating the chatbot's conversational role from its subsequent evaluative function. A system intended for language learning must distinguish between an error actually produced by the learner and one originating from speech recognition, model interpretation or conversational management. This becomes especially important

when the same system both participates in the interaction and subsequently evaluates it. The preliminary architecture proposed for PragmAI, which separates a conversational agent from a Tutor agent, responds directly to this limitation: the former would manage role, task progression, and situated interaction, while the latter would analyse the completed transcript according to explicit pedagogical criteria.

The findings shift the design question from how to write a better prompt to how to build an architecture that can activate pragmatic, procedural, and pedagogical knowledge in a controlled way. Prompt engineering and structured knowledge remain useful, since the experiments clearly show that they improve several aspects of interaction. However, the Custom GPT's residual instability suggests that prompting alone may be insufficient for a system expected to provide predictable pedagogical behaviour.

The next stage of the broader project therefore consists in investigating an agentic architecture in which conversational state, scenario-specific knowledge, pragmatic decision-making and pedagogical evaluation can be more clearly separated and coordinated. In this sense, the experiments reported here are useful not only as chatbot tests, but as a requirements-analysis phase for the subsequent development of PragmAI.

This design direction can also be situated in relation to recent studies exploring the pedagogical use of GenAI-based tools and chatbots for language learning. Mezzadri and Tangorra [14] examine integrating ChatGPT to develop writing skills, while Filippetti et al. [15] investigate learners' perceptions of a CEFR-based chatbot for Italian L2. The present study addresses a different issue: how an LLM-based conversational system can be configured to sustain task-based interaction in which pragmatic appropriateness, conversational progression, and pedagogical feedback must be managed simultaneously. Its contribution therefore lies in identifying the limits of prompt and knowledge-based customisation for such situated interactions and using them to inform the design of a more structured agentic system.

5. Conclusion

The preliminary experiments show that increasing prompt specificity and introducing structured knowledge can improve several aspects of an LLM-based chatbot for L2 pragmatics, including role maintenance, register, task progression, and feedback. At the same time, the tests also confirm that these improvements are not always stable or directly attributable to a single design choice. Across versions, several variables changed simultaneously, including the type and amount of instruction, the knowledge sources provided to the model, the interaction modality, the learner proficiency level, and the organisation of the configuration files. The study should therefore be interpreted primarily as an exploratory and requirements-oriented phase, useful for identifying recurrent critical areas and progressively refining the system, rather than as a controlled comparison between chatbot configurations.

A more systematic experimental phase is therefore required, with more precise control of configuration settings, experimental parameters such as the LLM used, and relevant learner characteristics, using a controlled procedure in which one variable is modified at a time, while the remaining conditions are kept constant. This will make it possible to investigate, for example, the independent effects of oral versus written interaction, learner CEFR level, prompt structure, the amount and type of contextual knowledge, and configuration architecture. Comparable interaction scenarios and evaluation criteria will also be used across conditions, allowing changes in chatbot behaviour to be associated more reliably with specific design modifications. A parallel comparison with other customizable LLM-based environments, including Gemini Gems, is also planned in order to examine whether the observed limitations are specific to the current platform or recur across different general-purpose conversational systems.

The next design step is to transition from complex Custom GPT configurations to a more controllable agentic prototype. The experiments reported here suggest that the main difficulty does not lie solely in providing the model with sufficient pragmatic knowledge, but mainly in ensuring that different functions are activated reliably and at the appropriate point in the interaction. The future PragmAI prototype will therefore explore a more explicit separation of functions, including conversational role management, task-state control, context-specific knowledge retrieval, and pedagogical evaluation. In particular, the distinction already identified between a conversational component and a Tutor component is expected to provide greater control over the transition from role-play to feedback and over attributing learner errors.

Future experimentation will test the agentic prototype under more controlled conditions before moving towards learner-based validation. This subsequent phase will make it possible to assess whether the

system can maintain a pragmatically coherent task-based interaction, as well as whether it can provide sufficiently stable, transparent and pedagogically reliable support for the development of Italian L2 pragmatic competence. In this sense, the preliminary tests described in this paper represent a necessary step from exploratory chatbot customisation towards the systematic design and evaluation of a specialised AI-assisted language-learning environment.

6. Acknowledgements

ChatGPT (OpenAI) was utilised as an artificial intelligence support tool for linguistic, stylistic, and editorial revision of the manuscript, as well as for reformulating selected passages. The authors remain fully responsible for the content of the paper, including the data, the analysis of the interactions, the interpretations, and the methodological and design choices presented.

REFERENCES

[1] Gesuato S., Castello E., “Pragmatics at university level? A survey of Italian EFL students’ perceived instructional experience and learning goals”, Pragmatic aspects of L2 communication: from awareness through description to assessment, 3, Newcastle upon Tyne, AA.VV., Cambridge Scholars Publishing, 2022, 38-78, ISBN 9781527577695, https://hdl.handle.net/11577/3416022.

[2] González-Lloret M., “Technology and L2 Pragmatics Learning”. Annual Review of Applied Linguistics, 39, 2019, 113-127, https://doi.org/10.1017/S0267190519000047.

[3] Long M., “Second Language Acquisition and Task-Based Language Teaching”, Malden, Wiley-Blackwell, 2015.

[4] Hubbard P., “General Introduction”, Computer Assisted Language Learning: Critical Concepts in Linguistics, I, London-New York, Routledge, 2009.

[5] Council of Europe, “Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume”, Strasbourg, Council of Europe Publishing, 2020.

[6] Bahari A., Han F., Strzelecki A., “Integrating CALL and AIALL for an interactive pedagogical model of language learning”, Education and Information Technologies, 30, 2025, 14305–14333, https://doi.org/10.1007/s10639-025-13388-w.

[7] García de Toro C., “Traducir la oralidad: su incidencia en el proceso de aprendizaje de la traducción”, Glosas Didácticas. Revista Electrónica Internacional, 13, 2004, 115–127.

[8] Escandell Vidal M. V., “Aportaciones de la Pragmática”, Enseñar español como segunda lengua (L2)/lengua extranjera (LE). Vademécum para la formación de profesores, Madrid, SGEL, 2004, 179–197.

[9] Ellis R., “Task-based language learning and teaching”, Oxford, Oxford University Press, 2003.

[10] Bhandari L. P., Dahal N., Awasthi J. R., Dhungana S., “Technology-mediated task-based language teaching: a systematic review”, Cogent Education, 12(1), 2025, https://doi.org/10.1080/2331186X.2025.2560051.

[11] Acharya D. B., Kuppan K., Divya B., “Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey”, IEEE Access, 13, 2025, 18912-18936, https://doi.org/10.1109/ACCESS.2025.3532853.

[12] Gruber J., “Markdown: Syntax”, The Daring Fireball Company LLC, available at https://daringfireball.net/projects/markdown/syntax (accessed: 22 September 2026).

[13] Briz Gómez A., “El análisis de un texto oral coloquial”, ¿Cómo se comenta un texto coloquial?, Barcelona, Ariel, 2000, 29–50.

[14] Mezzadri M., Tangorra M., “Integrare l’Intelligenza Artificiale nell’educazione linguistica: uno studio di caso con ChatGPT per lo sviluppo delle abilità di scrittura”, Italiano LinguaDue, 17(2), 2025, 639-658, https://doi.org/10.54103/2037-3597/30458.

[15] Filippetti S., Cinganotto L., Montanucci G., “AIDI, a CEFR-based chatbot for Italian as a second language: pilot survey on learners’ perceptions”, Advancing CALL: New research agendas - EUROCALL 2025 Short Papers, 2025, https://doi.org/10.4995/EuroCALL2025.2025.21283.