Umanistica Digitale – ISSN: 2532-8816 – n. 16, 2023
S. Lilli – *ChatGPT-4 and Italian Dialects: Assessing Linguistic Competence*
DOI: http://doi.org/10.6092/issn.2532-8816/18221

ChatGPT-4 and Italian Dialects: Assessing Linguistic Competence

Silvia Lilli

Department of Literary Studies, Philosophical Studies and History of Art, University of Rome "Tor Vergata", Rome, Italy
silvialilli@hotmail.it

Abstract

The purpose of this study is to evaluate ChatGPT-4's language proficiency in Italian dialects. At the outset, it is clarified what is meant by 'language ability' within the context of Large Language Models. This involves identifying the tasks that ChatGPT might face in real-world scenarios, from which we can derive inferential assumptions regarding its linguistic ability. The skills identified, which served as foundation for test design, include comprehension and translation, dialect recognition, analysis of the distinctive features, error detection, text production, interaction, theoretical background, and self-assessment. The tests were crafted to mimic situations requiring these competencies, trying to emulate authentic ChatGPT-User interactions. The results highlight ChatGPT's excellent prowess in understanding and recognizing Italian dialects and their subvarieties, and a robust background and awareness of its own knowledge. However, the model exhibits significant gaps in analytical skills and struggles with text production and interactive tasks, suggesting superior passive linguistic capabilities compared to active ones.

Keywords: ChatGPT, LLMs, Italian Dialects, Language Testing, Language Abilities, Machine Translation

Questo studio si pone come obiettivo la valutazione della competenza linguistica di ChatGPT-4 in relazione ai dialetti italiani. Per prima cosa, viene chiarito cosa si intende per 'abilità linguistica' nel contesto dei Large Language Models: ciò significa identificare concretamente le attività in cui ChatGPT potrebbe essere coinvolto in situazioni reali, dalle quali derivare conclusioni inferenziale sulle sue abilità linguistiche. Le competenze individuate, alla base dei test di valutazione elaborati, sono le seguenti: comprensione e traduzione, riconoscimento dei dialetti, analisi delle caratteristiche distintive, individuazione degli errori, produzione testuale, interazione, conoscenze teoriche e autovalutazione. I test sono stati elaborati cercando di riprodurre situazioni reali che richiedono l'impiego di queste abilità, simulando un'interazione autentica tra ChatGPT e utente. I risultati hanno rivelato un'eccellente capacità da parte di ChatGPT nella comprensione e nel riconoscimento dei dialetti italiani e delle loro varianti, così come un solido background teorico e una buona consapevolezza delle proprie conoscenze. D'altra parte, il modello presenta importanti lacune nell'analisi linguistica e difficoltà nella produzione di testi e nell'interazione conversazionale, suggerendo una maggiore attitudine per le capacità linguistiche passive rispetto a quelle attive.

Parole chiave: ChatGPT, LLMs, Dialetti Italiani, Valutazione linguistica, Abilità linguistica, Machine Translation

Introduction

To our knowledge, ChatGPT has been trained on a vast array of written languages from around the world, acquiring the ability to comprehend a broad range of linguistic material. The precise mechanism by which ChatGPT has learned even less prevalent languages, presumably benefiting from its training on more well-represented languages, remains somewhat undisclosed. This study emerged, thus, from a singular question: does ChatGPT 4.0 understand Italian dialects too? To address this question, various tests have been tailored, which will be discussed in subsequent sections. Despite the selective choice of these and of the dialects tested, this study is intended to be a preliminary exploration in this field, aiming to foster a deeper understanding of the AI system's capabilities and of the boundaries of its extensive knowledge base.

State of the art

Large Language Models' (LLMs) capacity for human-like communication has significantly transformed the interface between humans and machines, offering a more natural and appealing form of interaction. Across them, OpenAI's ChatGPT, a chatbot based on multimodal LLMs, has affirmed its prominent position, especially its 4th version released in March 2023 [38]. Its recognized excellence in Natural Language Processing (NLP) tasks has certainly contributed to its success [31], raising numerous questions about its potential application to a wide range of different practical tasks and real-world scenarios [24]. Much research has been conducted in the past months to explore the extent of the model's abilities, its boundaries, and related ethical concerns [44], [46], [52], [58], [53]. Many of them tried to assess ChatGPT's performance on specific NLP tasks, problem solving [57], information extraction [25], [33], machine translation [27], confronting test results with state-of-the-art benchmarks, underscoring higher and lower performance [30], and evaluating its errors for improvement [9].

Similarly, this contribution aims to explore ChatGPT-4's behaviour on the specific linguistic domain of Italian dialects. Comparable research has been conducted for other well-represented natural languages, such as Chinese [34], as well as for lesser-represented languages [32], sometimes showing results inferior to those of existing tools, or exhibiting significant variance in performance across different languages. For this research, I will not refer to an existing dataset or a standard benchmark testing set. The tests were designed from a humanistic perspective addressing an audience of humanists who are not specialists in Computer Science. Thus, the use of coding or mathematical scoring procedures is minimized. On one hand, this may face criticism due to potential personal judgment biases and the limitations of the tests compared to broader datasets used in automated testing methodologies. Nevertheless, allowing humanists to investigate LLM's performance with their own methods could bring new insights in this research area, encouraging a closer collaboration between Humanities scholars and specialists in Computer Science.

Target Dialects selection

The first step in this study is to select the Italian dialects that will serve as samples for tests. This operation turns out to be difficult, given the wide spectrum of Italian dialects and their significant

variation in terms of linguistic features, distribution, prestige, and persistence. A comprehensive study should, indeed, include a broad sample of dialects from distinct categories. According to the prevalent classification, Italian dialects are categorized based on their linguistic similarities into four main groups, as firmly defined by Pellegrini in 1975 [41]: Northern, Friulian, Middle, Southern, and Sardinian. This classification evidently relies on linguistic and, consequently, geographic criteria. The aim of this study, however, is not to establish a descriptive system, but to detect potential disparities in ChatGPT's knowledge extent. Such disparities could be influenced by two factors: the degree of linguistic divergence in relation to a better-known reference language (Italian)[1] – under the hypothesis that a dialect more similar to Italian may be better understood – and the distribution[2] of the dialect, which impacts the breadth of materials that might have been part of ChatGPT's training. By combining the two criteria – linguistic deviation from the standard language and the spread of dialect usage – we derive the following dialect classification:

1) Widely distributed, highly characterized (e.g., Sicilian, Venetian).
2) Widely distributed, minimally characterized (e.g., Roman, Tuscan).
3) Narrowly distributed, highly characterized (e.g., Genoese, Sardinian).
4) Narrowly distributed, minimally characterized (e.g., Abruzzese, Umbrian).

This classification, being purely theoretical, is open to debate regarding the exact placement of a dialect within each group. Dialects, moreover, exist along a geographical and linguistic continuum rather than strictly adhering to defined categories. Nonetheless, this classification underscores the limitations of research that only considers dialects from a single category, suggesting that results can significantly vary across these parameters of variation.[3]

In this case study, for reasons of time and resources, I focused only on three dialects from the first two groups: Sicilian and Lombard for the first group, and Roman for the second.[4] This

[1] Other attempts at classifying Italian dialects, including the pioneering work by Graziadio Isaia Ascoli [2], are based on measuring the linguistic distance between the dialectal variety and standard Italian (i.e., the Tuscany variety), thereby emphasizing the role of the reference language, much as I do.

[2] In this context, "distribution" refers primarily to the number of speakers, but also to dialect's prestige, which may be derived from its cultural (literary, theatrical, cinematographic etc.) or political tradition of usage.

[3] In this paper, I will use the terms 'language' and 'dialect' somewhat interchangeably to refer to diatopic varieties of language within the Italian territory. This choice is rooted in the historical development of Italian dialects from Latin, which, from a linguistic standpoint, qualifies them as distinct 'languages'. However, from a sociolinguistic perspective, their relationship with Italian as a *Dachsprache* (umbrella-language) categorizes them as dialects [35]: 3-32; [3].

[4] More precisely, I will focus on a communal variety for both the Roman dialect and Lombard, specifically the Milanese variant, while adopting a regional variety for Sicilian. Although Sicilian has substantial differences across its subvarieties [47], [55], [5], the published text chosen for this analysis (detailed below) does not specify a particular subvariety. Therefore, Sicilian will be treated as a singular dialect in this study, although I acknowledge the possibility of more detailed differentiation in subsequent studies. The Roman dialect is predominantly spoken in the capital city, distinguishing it from the regional surroundings with their distinct linguistic nuances, even if its status as the capital's variety places it in a privileged position, extending its influence [54], [18], [4]. Milanese, on the other hand, might be viewed as a regional dialect due to the pronounced

selection was based on the following criteria: availability of reliable texts for each dialect; personal familiarity with the dialects; a desire to represent at least a broad geographical spectrum (North, Central, South) of dialect variations. However, the results should be interpreted within the context of these limitations and further research is desirable for a comprehensive understanding of ChatGPT's proficiency in Italian dialects.

Language ability and test designing

Before discussing the testing procedures, it is necessary to explain what is meant by 'language ability' and how it can be effectively assessed. This topic lays the foundation for this research approach and is particularly sensitive, as it delves into theoretical questions that have been the subject of debates for decades [13], [56], [6]: 61-82.

Historically, the debates have been conducted trying to define and assess human linguistic competence. Hence, it is worth noting that in this experiment I will transpose assumptions made for human behaviours onto machine behaviours. Rather than on similarity of cognitive process, this transposition is grounded in the apparent similarities in linguistic responses between ChatGPT and human speakers, which position ChatGPT as a potential surrogate for humans in many linguistic tasks, leading to the emergent implications we observe. Yet a pivotal question arises: Can the assumptions we make about human speakers truly be applied to LLMs, especially considering that LLMs neither understand nor retain linguistic competence in the same manner as humans? This is a crucial and delicate question, and it is confronted with the limitations in our current understanding of both human brain and Artificial Intelligence functioning. For the purpose of this study, while acknowledging the question's importance, I will not address it directly, proceeding under the premise that the similarity in outputs between humans and ChatGPT allows us to apply similar methods in investigating its functioning. Nevertheless, a deeper exploration of this topic and wider methodological discussions are strongly desired for solidifying the foundations of any research in this domain.

Following Bachman and Palmer [6]: 66-78, 'language ability' can be defined as the convergence of two elements: language knowledge and strategic competence. Here, strategic competence refers to the metacognitive strategies that allow an individual to perform a specific linguistic task. In this light, language use is not an abstract concept but is always a performance realized in a specific situated task. Therefore, when evaluating language ability, the traditional four language skills (reading, writing, speaking, listening) appear inadequate. Instead, it is the specific testing situation that should be considered, which encompasses the test's purpose, the test-takers, and the Target Language Use (TLU) domain. As they note, "the way we define language ability for a particular testing situation, then, becomes the basis for the kinds of inferences we can make from the test performance" [6]: 66.

Reflecting on these considerations, our initial step is to define the TLU in our context, i.e., "a set of specific language use tasks that test taker is likely to encounter outside the test itself, and to which we want our inferences about language ability to generalize" [6]: 44. In other words, before designing a language assessment test, we should pose the question: What are the situations in which ChatGPT might engage with dialects in its 'real-world' domain? This means taking into consideration the distinctive characteristics of the test subject (artificial rather than human),

influence of the city centre on nearby areas, or, at the very least, it serves a representative of the Wester Lombard subvariety [36], [48], [49], [10].

ChatGPT's specific interactional environment (chatbot) and its primary designed task (text production). I outlined a list of specific tasks which may involve dialect use in ChatGPT-User interaction:

- Reading and understanding an input text in a target dialect. (Ability: Understanding).
- Recognizing a specific dialect and distinguish it from others. (Ability: Discriminating).
- Analysing a text in a specific dialect, pinpointing specific linguistic features, or detecting errors (Ability: Analysing and Correcting).
- Coherently interacting using the language domain of a given input. (Ability: Interacting).
- Producing a text using the target dialect (Ability: Text Producing).
- Showcasing a comprehensive theoretical knowledge of the linguistic and extralinguistic features of the target dialect (Ability: Language Knowledge).
- Being aware of the extent and the boundaries of its knowledge and evaluating it (Ability: Self-Assessment).

Starting from these tasks that could occur in non-test situations, the next phase involves designing tests that reproduce the listed abilities. These tests, in addition, should be guided by a model of test usefulness which encompasses qualities such as reliability, construct validity, authenticity, interactiveness, impact, and practicality.[5] This model is largely subjective, as it depends on developer's values judgment, but it still should rely on a validation framework that can guide through all the phases of the test construction.[6]

I began by stating the assumption that ChatGPT's proficiency in a particular dialect, contextualized within the interaction between User and Language Model, implies the aforementioned abilities. The testing procedures are then structured into four distinct parts as delineated below:

A. Comprehension and Translation. This section assesses the ability to comprehend the language, through two different strategies: 1) Translating from the target dialect to a better-known language; 2) Answering comprehension questions based on the provided

[5] According to Bachmann and Palmer [6], [7], we can summarize each quality as follows. Reliability: consistency of measurement; Construct validity: meaningfulness and appropriateness of the interpretations that we make on the basis of test scores; Authenticity: correspondence between characteristics of TLU and characteristics of the test task; Interactiveness: the ways in which the test taker's area of language knowledge, metacognitive strategies, topical knowledge and affective schemata are engaged by the test task; Impact: the effects that decisions based on test evaluation have on test-takers life and to social contexts; Practicality: the rapport between the available resources and the required resources. Not all these qualities apply to our contexts, such as impact or affective schemata.

[6] A framework, as firstly delineated by Kane [28], [29], that can essentially be summarized in four steps: giving scores to single observation (*scoring*), using them to generate an overall score representing performance in test setting (*generalization*), drawing an inference regarding what test score might imply for real-life performance (*extrapolation*) and then interpreting this information and making a decision (*implications*).

text. The underlying idea is that word-to-word translation, as well as rephrasing and summarizing, require solid linguistic knowledge. The second strategy also examines the inferential ability required to extrapolate explicit and implicit information from the dialectal text.

B. Discrimination and Analysis. This section evaluates the abilities to discriminate, analyse, and correct errors across three different phases. In the first two phases, the task performance strategy involves identifying the linguistic features that differentiate various language varieties. In the first task the recognition process is implicit (merely recognizing the dialect), while in the second task it becomes explicit (indicate the dialect distinctive features): this is meant to demonstrate an awareness of the process accomplished in part one. The analysing ability is also assessed in part three, where ChatGPT is asked to detect errors and provide corrections.

C. Interaction and Translation. This section evaluates the abilities to interact and produce texts. For interaction, the test simulates a conversation in a specific dialect, assessing the ability to respond coherently within the language domain and maintain consistent use of the dialect throughout the conversation. For text production, a second translation task has been designed (from a well-known language to the target dialect), based on the premise that proficient translation implies text production skills.[7]

D. Theoretical Background and Self-Assessment. The last section assesses ChatGPT's possession of a solid theoretical language knowledge. The second part of the test, indeed, involves the fundamental metacognitive strategy of assessing the test-taker's own knowledge and mastery. Both parts consist of a simple question-and-answer format. This method is chosen because it directly addresses the aspects of consciousness and awareness that the test aims to measure.

This list represents just one of the potential effective methodologies for assessing linguistic proficiency and can certainly be strengthened with additional criteria and tests. Nonetheless, it tries to encompass a diverse set of abilities and strategies to provide a comprehensive overview of ChatGPT's dialectal competencies.

The four parts of the test designed are thus articulated in a series of detailed procedures as follows:

A. Comprehension and Translation (from dialect)

[7] I could have chosen to explicitly request ChatGPT to produce a new text (whether fiction or non-fiction) on a specific topic, and then evaluate its linguistic accuracy. However, I opted for a counter translation task because the availability of a reference translation and a predetermined scoring system simplified the assessment process. Assessing the language accuracy of an original text would have necessitated the involvement of other individuals (especially for Milanese and Sicilian, which are not my native dialects) and would have relied on a more subjective evaluation rubric. A general evaluation of less constrained production skills has indeed been already tested in C.1.

1. Submit a text in dialect and request a translation in a target language. Evaluate the accuracy of the translation according to predefined assessment criteria.
2. Ask some questions that require the comprehension of the text to answer.

B. Discrimination and Analysis

1. Submit a text in dialect and ask to identify the type of dialect and its geographical distribution. Proceed with dialects with decreasingly widespread distribution.
2. Submit two texts in two varieties of the same dialect group and request ChatGPT to distinguish them and indicate their respective geographical distribution; then request ChatGPT to analyse the texts, identifying the distinctive linguistic features that facilitate the differentiation.
3. Present dialect texts containing intentional grammatical errors. Request ChatGPT to detect these errors, identify their nature, and provide appropriate corrections.

C. Interaction and Translation (to dialect)

1. Begin a conversation in a specific dialect. Test if ChatGPT can respond coherently in the chosen dialect, comprehend the context, and offer consistent replies.
2. Submit a text in a well-known language to ChatGPT and ask for translation in each dialect of the set. Measure the accuracy of translations using predetermined evaluation criteria.

D. Theoretical Background and Self-Assessment

1. Ask ChatGPT to explain the characteristics of each dialect of the set, describe its distribution, and its orthographic conventions. Then Ask ChatGPT to explain the sources of its knowledge of Italian dialects, and to evaluate it. Finally, ask ChatGPT how its knowledge can be improved in this field.

Testing procedures and issues

In this chapter, I detail the testing procedures and discuss the challenges encountered during test design. Before delving into the topic, it's important to highlight that all tests were conducted using a Zero-Shot approach. This means submitting the prompts without providing any examples of the expected procedure or outcome [11]. While some may argue that Few-Shot In-Context-Learning or Chain-of-Thought approach may yield better results,[8] my intention was to simulate a natural interaction between a user and ChatGPT. For brevity in this document, I've reported only the specific prompts used and their resulting scores. Comprehensive findings will

[8] See Peng [42] for discussion and further references.

be detailed in Results section. The full versions of all texts, scripts, outcomes, and result sheets referenced in this research can be accessed in a dedicated GitHub repository.

Comprehension and Translation (Tests A.1 and A.2)

The objective of these tests is to evaluate ChatGPT's ability in understanding the language presented. Since the first test requires translation from a source text in dialect to another target language, a primary challenge was sourcing reliable texts in the three selected dialects. The chosen texts needed to be reliable in representing each dialect, coherently aligning in terms of length, difficulty, chronology, and genre. Additionally, I aimed to use texts unfamiliar to ChatGPT, which meant avoiding sources that were surely part of the training data, like literary works or web pages (such as Wikipedia dialectal editions).[9] This strategy was intended to reduce potential biases or misrepresentations in the results. I finally opted to use an extract of Chapter 8 taken from *Le Petit Prince* by Antoine de Saint-Exupéry, which has been translated in a vast array of languages, including several Italian dialects: although it is uncertain to what extent literary works in Italian dialects have been included in the training data, the fact that these texts are only available in print editions is a supporting argument. Moreover, the parallel nature of the translations provides an opportunity for a consistent comparison across the dialects, and using a published text ensures a degree of accuracy and coherence in the material.[10] Another determining factor is the easy availability of these translations [22], [20], [21] in online bookstores.

Once defined the source texts, the second question regarded the choice of the target language. It seemed more effective to choose Italian, as dialects tend to have a linguistic domain closer to Italian, even if they developed autonomously from Latin and many of them possess the status of 'language'. Nowadays, in fact, dialects are diatopic varieties of Italian language, and it is likely that a user interacting in dialect would probably be an Italian speaker.[11]

To evaluate the accuracy of translations, both for points A.1 and C.2, I could leverage indexes already existing for Machine Translations (MT). Specifically, three different measures were used for scoring: the BLEU score [39], the TER score [51], both of which confront a source translation with a reference translation, and the COMET score [45],[12] which additionally

[9] The Wikipedia edition in Lombard was initiated in 2005 and hosts today (September 2023) 73,349 articles and 145,222 pages. The Sicilian edition began in 2004 and currently has 26,256 articles and 55,967 pages. Notably, there is no Wikipedia edition for the Roman dialect, but there is a discussion on this topic among Wikipedia Community's members.

[10] This encompasses linguistic features as well as graphical conventions. Dialects often lack rigid and widely accepted grammatical and orthographical rules, and many variations can be attributed to diatopic differences, particularly in the absence of a unifying political centre. However, the existence of a strong literary tradition for the three selected dialects partially alleviates some of these concerns. In this case, the translations of *Le Petit Prince* are done under the supervision of local scientific academies, directly engaged in the study and preservation of their dialectal heritage.

[11] Italian, thus, is also used as source language for translation test C.2.

[12] The BLEU (Bilingual Evaluation Understudy) score evaluates machine translations based on how many n-grams they share with reference translations. TER (Translation Edit Rate) quantifies the number of edits required to change a system output into one of the references. COMET is a modern evaluation metric that uses neural networks to predict human judgments of translation quality, considering various aspects beyond n-grams. All these metrics are implemented for

confronts with the original text.[13] However, for tasks beyond translation – open-ended questions, error detection, dialect recognition, conversation simulation – I chose not to use an automatic scoring system based on NLP and Machine Learning techniques. Instead, I employed a human-centric, rubric-based scoring approach, eventually supported by a checklist to enhance the objectivity of observations. Indeed, human scoring rubrics offer flexibility to adapt to task requirements, while automated scoring systems often necessitate substantial modifications or retraining. Moreover, their appropriateness becomes particularly questionable with smaller datasets like this and especially for tasks marked by high subjectivity or without a definitive 'correct' response.[14] A description of different evaluation systems used in each test will be detailed in the Results section.

Once defined the target language, another issue was to select a reference translation for confrontation. The initial hypothesis was to use one already existing Italian translation of the text, assuming it served as the basis for dialectal versions. However, relying on it presented the risk of comparing two independent translations from the original French, which potentially possess more variations than expected, leading to an underestimation of ChatGPT's performance. Also, multiple translations exist from French to Italian, making it difficult to pinpoint which specific Italian translation the dialectal versions might have been derived from (but on this point see note 23). Secondly, the primary objective of this test is to assess the comprehension of the given language, not necessarily the quality of the translation itself. Therefore, a literary translation might introduce various stylistic elements that could widen the gap between the two texts, and that respond to expressiveness' needs not directly connected to the translation procedure.

For all these reasons, the most pragmatic and robust approach appeared to derive a new translation directly from each dialectal text. This translation, undertaken by the author with the support of dialectal dictionaries,[15] responded to the utmost effort of retaining the original phrasal

Python environment within NLTK package [8]. For an interesting implementation of COMET to language dialects (including Italian ones) see Alam et al. [1].

[13] Another interesting test could evaluate the different responses in a Task Specific Prompt setting. I didn't investigate it in a systematic manner, given that few random experiments showcased an irrelevant variation in the results. Anyway, it cannot be excluded that more extensive research could highlights substantial differences, as showed in the case of other languages [42], [27].

[14] As highlighted by the reviewers of this paper, I acknowledge that the use of an author-designed scoring system and the author's sole scoring may introduce a degree of subjectivity into the evaluation process. To achieve more solid statistical results, it would be recommended to involve multiple raters and establish a baseline through Inter-Rater Reliability (IRR) scores. Despite this, my research is intended as an exploratory investigation, and even in a non-strictly experimental setting it can yield some useful insights into ChatGPT's functioning. Several studies, indeed, are following a similar strategy: see, for instance, the recent studies on ChatGPT's abductive reasoning [40], or Microsoft Research early experiments on ChatGPT's potential for Artificial General Intelligence [12].

[15] For Sicilian-Italian, the best dictionary is edited by Piccitto [43]; for Milanese-Italian, apart from the classic Cherubini's [14], is an easy but valid edition the recent *Dizionario Milanese* edited by the Circolo Filologico Milanese [15]; regarding the Roman dialect, there is still a lack of scientific works. This void is being addressed by the edition of the *Vocabolario del Romanesco Contemporaneo*. However, as of now, only the first two volumes covering letters I, J, and B have been published [17].

structure and the precise meaning of individual terms. To ensure that ChatGPT would also attempt a literal translation, the temperature parameter must also be taken into consideration. The temperature parameter influences the randomness of ChatGPT's responses; a lower value produces more deterministic outputs, while a higher value allows for greater variability. Consequently, and with the intent to verify its effects on translation task [42], the tests were carried out three times: the first, without any indication about temperature's setting, then requesting ChatGPT to provide translations under two distinct temperature parameters, 0.2 and 0.6.[16]

Below are the prompts meeting the specified criteria.[17] A brief contextualization is provided at the outset. The instruction to not recognize the source aims to prevent ChatGPT from referring to the literary source, considering a possible familiarity with such a renowned text.[18] The prompts that were not included in first attempt (with default temperature setting) are enclosed between square brackets.

I'd like to test your proficiency in Italian dialects. I'll provide a short text in one of the Italian dialects. Proceed with the following steps: 1) Recognize the dialect in which the text is written. 2) Without recognizing the source of the text, translate the text from dialect to standard Italian, [setting your temperature at 0.2. 3) Without recognizing the source of the text, translate the text from dialect to standard Italian, setting your temperature at 0.6. 4) Only after having translated the text, try to recognize the source from whence the text has been taken.] Input text: "…"

Table 1. Test A.1 prompts.

[16] ChatGPT, indeed, cannot modify its parameter settings on its own; adjustments to temperature and other parameters can only be made through the ChatGPT Application Programming Interface (API). Nevertheless, ChatGPT understands that a value close to 0 leads to more deterministic responses, while a value near 1 results in more imaginative outcomes. This awareness thus influences the system's outputs. However, it is important to note that the values '0.2' and '0.6' should not be perceived as the exact parameter set.

[17] The prompts, here and in all the tests, are presented in Italian, given, as already noted, the close relation between dialects and the Italian linguistic context. Here only the English translation is reported; for the original Italian prompts see the GitHub repository. For prompts designing see Gao et al. [23].

[18] It should be noted that such a prompt might not work in a strictly literal sense, as ChatGPT cannot intentionally exclude specific knowledge from its outputs. The prompt was inserted because, during preliminary trials, ChatGPT quickly recognized and commented on the source. Therefore, the prompt aimed to steer ChatGPT toward the translation task rather than delving into literary considerations.

Regarding test A.2, the purpose of the questionnaire was to assess the effective comprehension of the text submitted. Again, the questions are presented in Italian and imply not only the capability of reading the test (i.e., questions whose answer are incorporated within the text), but also to understand the meaning in a deeper manner (i.e., questions whose answer require an inferential process). The questionnaire has been submitted twice: before and after the translating operation, to verify possible differences in the behaviour depending on the resort to the translated text as support for comprehension. The intention was to point out eventual misunderstanding, and to assess the quality of the machine understanding.

Below are the prompts submitted (the sentences in square brackets were included in the questions posed before the translation process).

[Now I will present you with a text in an Italian dialect. I would like you to read the text and, without translating it, answer some comprehension questions. Here is the text: "…"]
Now I would like to ask you some questions about the text, to test your understanding.
Answer each question in order:
1) Why is the little prince surprised by the new sprout?
2) Why does the flower claim to be wrinkled?
3) Why does the flower consider itself beautiful?
4) Why is the flower touching according to the little prince?
5) What gesture is requested by the flower from the little prince?

Table 2. Test A.2 prompts.

Discrimination and Analysis (Tests B.1, B.2 and B.3)

This phase of the test was designed to assess more analytical competencies. Tasks B.1 and B.2, which involve recognizing a text's dialect and distinguishing between two similar varieties, necessitate the ability to synthetize and attribute specific linguistic characteristics to a particular dialect within ChatGPT's knowledge base. This becomes particularly challenging when dealing with a dialect that lacks a broad literary tradition and shares substantial linguistic similarities with a more prominent variant.

As will be evidenced in the responses to the prompts in section A.1, question 1 (see below), ChatGPT showed no hesitation in identifying the three widespread dialects chosen for this study. Therefore, the testing proceeded with less prevalent varieties or subvarieties of a regional dialect. The texts selected for analysis were in Bolognese, Calabrian and Friulian dialect, each ranging from approximately 120 to 190 words in length, two in prose (Bolognese and Friulian), one in poetry (Calabrian). As these texts were sourced from various web pages, there is no guarantee of accuracy in terms of linguistic and orthographic conventions, despite efforts to prioritize reliable sources.[19] It is important to note, moreover, that utilizing only one text for each dialect, as done here, does not ensure statistically significant conclusions. Set of tests including different

[19] While translated version of *Le Petit Prince* may have been available also in these specific dialects (or others), I opted to conduct tests with readily accessible texts due to the constraints of time and budget. This decision aligns with the goal of the tests, which was just to evaluate ChatGPT's ability to identify dialects without translating it.

typologies of texts would be required for that purpose. However, as we will see in Results section, the stability of the responses (successful for the recognizing task, unsuccessful for the analytical task) permits to drive tentative assumptions.

For test B.2, the comparison was drawn firstly between the Bolognese text and an excerpt in Modenese dialect, the first representative of the Western group of Emilian dialects, the second of the Eastern. The next comparison was between a text in Bergamo's dialect and a segment of the Milanese version of *Le Petit Prince* used in test A. Both belong to the Lombard dialect group, with the former being representative of the Eastern and the latter of the Western subvariety. This phase began with a request to determine the potential identity of the two dialects, followed by an analytical task aimed at identifying the distinguishing features leading to such conclusion.

Below are the sources for each respective text, followed by the prompts submitted for each task.

- Emilian (Bolognese): Marchetti, Gaetano. 1968. *Ai ténp dal pôver Scarabèl*.
- Calabrian: Pelaggi, Bruno. 1880. *A 'Mbertu Primu*.
- Friulian: Regione Friuli-Venezia Giulia, 2007. *Regolament pe concession dai contribûts pe promozion de lenghe furlane*.
- Emilian (Modenese): Società del Sandrone. 2022. *Sproloquio della famiglia Pavironica*.
- Bergamo dialect: Mastrocco, Giorgio, ed. 2019. The Italian Constitution translated in Bergamo Dialect.

Task 1: I would like to test your competence in recognizing Italian dialects. I will submit to you as input a text in one of the Italian dialects. Read the text and perform the following operations: 1) Recognize the dialect in which the text is written. 2) Indicate the geographical area where this dialect is spoken.
Task 2: I would like to test your competence in recognizing Italian dialects. I will submit to you as input two texts in two different Italian dialects. Read the texts and perform the following operations: 1) Identify whether the texts are written in the same dialect or in two different dialects. If they are different dialects: 2) Identify the area where the dialects are spoken. 3) Identify the linguistic differences that allow distinguishing between the two dialects. Text 1 "…"

Text 2 "…"

Table 3. Test B.1 and B.2 prompts.

In test B.3, I deliberately introduced six grammatical errors into the texts already used in test A.1. The errors pertained to subject-verb or noun-adjective/article agreement (number and gender), incorrect verb forms (e.g., conditional and infinitive instead of the present tense), reduplication of clitics and omission of crucial sentence constituents such as verbs. These errors were designed to violate general grammar rules that are applicable both in dialects and in standard Italian, rather than rules specific to the dialects themselves. This approach was chosen because less formally structured languages often have a wide range of acceptable forms that reflect local influences and variations. Furthermore, I assumed that general grammar rules violations would be easier to detect compared to dialect-specific rules. I tried to introduce the same errors in similar positions across the three texts. However, variations in the texts and their distinct grammatical structures sometimes required slight adaptations.

Below is the list of the errors introduced for each dialect, followed by the correct word form, and the prompts submitted.

Sicilian	avia/avianu; accabbarìa/accabbau; cumpariria/cumpariri; cucche/cucca; aviri/avia; bedda/beddu.
Milanese	era/eran; dismettess/dismettuu; sortiva/sortì; eran/era; sperlusciada/sperlusciaa; omission of verb "dii".
Roman dialect	era/erano; smettesse/smise; compariva/comparì; fatti/fatto; spettinata/spettinato; vedellolo/vedello.

Table 4. Error inserted in dialectal text for test B.3.

I will submit a text in an Italian dialect to you. Proceed with the following operations: 1) Read the text. 2) Without translating the text, identify the presence of any grammatical or syntactic errors. 3) If you don't find grammatical or syntactic errors, return the phrase "I did not find any grammatical or syntactic errors". 4) If you find grammatical or syntactic errors, return the phrase "I found grammatical or syntactic errors, in the number of [n]", where [n] is equal to the number of errors you have identified. 5) If you have identified errors, list them one by one in a numbered list, quoting the portion of the sentence where the error is located, followed by the proposed correction for the identified error.

Table 5. Test B.3 prompts.

Interaction and Translation (Tests C.1 and C.2)

The intention of the test C is to assess ChatGPT's ability to converse in dialects such as in the other languages within its training data (C.1) and to translate from Italian to the specific dialect (C.2). It tests the productive use of the language for interactive real-life situations and text production.

In relation to test C.1, from the versions of *Le Petit Prince* utilized in test A, I derived a series of dialectal greetings and questions, slightly adjusted to encourage a conversational response from ChatGPT and to integrate feedback from its earlier interactions.[20] The aim in designing the questionnaire was to prevent distorted replies from ChatGPT, particularly those overly extended, as ChatGPT's capability to respond in dialects significantly diminishes with longer replies.[21] Hence, to provide inputs more aligned with the testing goals, the questions set was refined through multiple iterations, before being finally submitted for evaluation in a new conversation. I present only the English translation of the prompts here; the dialectal versions are available in the linked GitHub repository.

The test was conducted in two rounds. In the first the conversation in dialect was initiate by directly posing a question in dialect. Given this approach, to facilitate the detection of the conversation language, the initial question was prefaced with small comment to augment the linguistic material. In the second round the specific domain was delineated clearly by stating the language used for the inputs and instructing ChatGPT to adhere to it in its responses. This instruction was conveyed in Italian.

[I will start a conversation with you in <…> dialect, I would like you to reply to me in <…>.][22]

Good morning. Our planet is very beautiful. Where are you from?

- What is the purpose of being rich?
- What do you recommend I do?
- What does 'to meditate' mean?
- How does one meditate?
- What is an application?
- Whose are (the applications)?

[20] ChatGPT's responses slightly changed in the several attempts until the construction of a well-balanced dialogue, with the result that not always the interactiveness with the previous answer is maintained. However, this was considered not essentially given the non-human nature of the tester, and thus the lack of affective schemata that could shape the response.

[21] However, the concluding portion of the dialogue features questions pertaining to a more specialized field, specifically technology. This design choice aimed to assess the consistency of dialect usage even in contexts that are more formal and require descriptive tasks.

[22] The <...> is a placeholder for each specific Italian dialect; in square brackets is the instruction omitted in first attempt.

- Thank you. Good night.

Table 6. Test C.1 prompts.

For C.2, insights from test A.1 are applicable to the adopted methodology. I used the same text portion from test A for the counter-translation, inverting the source language from dialect to Italian and the target language from Italian to dialect. I tested two different Italian versions as sources: my translations derived from dialects (see above), and the renowned Italian translation by Nini Bompiani Bregoli [19]. This choice aims to assess the optimal input for ChatGPT. Using my literal translation might yield outputs more closely aligned with the reference. On the other hand, a more literary rendition might, conversely, encourage a more fluent translation, thereby closer to the published dialectal versions used here as reference.[23] The outcomes were evaluated using the same scoring metrics of test A.1: BLEU, TER, and COMET. Below are the prompts for the task.

I want to test your skills in the field of Italian dialects. I will provide you with a text in Italian. Without recognizing the source of the text, translate the text from Italian to the <...> dialect.

Table 7. Test C.2 prompts.

Theoretical Background and Self-Assessment (Test D.1)

These final tests were designed to evaluate ChatGPT's theoretical knowledge and self-assessment competences. There are two underlying questions. The first is: "Does ChatGPT possess solid linguistic knowledge in the field of Italian dialects?". We define 'linguistic knowledge' as the combination of topical and contextual information which, when paired with strategic competence, results in language ability, as described by Bachman [6]. The second question is trickier: "Is ChatGPT 'aware' of its own expertise and the boundaries of its proficiency in this field?". It is important to underscore the inherent limitation of this test: while a language model like ChatGPT houses a vast repository of knowledge, it isn't primarily designed for encyclopaedic recall, nor can we attribute to it such metacognitive qualities as consciousness or genuine reflection, even when the system simulates them [26]. Nonetheless, test outputs can provide interesting insights about the LLM's algorithmic processes which, in some ways, function analogously to human metacognition. Furthermore, these outputs can be insightful when comparing ChatGPT's self-assessment to the results of previous tests.

The test is structured as an open-answer questionnaire. Initial questions concentrate on a specific dialect, inquiring about its linguistic traits, geographical distribution, and potential orthographic conventions. The following questions ask ChatGPT to judge its own knowledge and capabilities in the area, identifying its information sources, and pinpointing areas that could be strengthened.

I'd like to ask you some questions about Italian dialects. I'd like to focus specifically on the <...> dialect. Please respond to the following questions: 1) In which geographical area is the <...> dialect spoken?

[23] Furthermore, it seems plausible that two of the dialectal versions derived from Brigoli's translation, as underscored, for instance, by the use of the terms «caffellatti» and «cappuccino» to render Brigoli's «caffè e latte», diverging from the original «petit déjeuner».

2) What are its linguistic features that differentiate it from standard Italian?
3) Are there any recognized orthographic conventions for this dialect?
4) What are the sources of your knowledge about the <...> dialect?
5) How would you assess the level of your knowledge about the <...> dialect?
6) Which aspects do you believe could be improved regarding your knowledge of this dialect, and how?

Table 8. Test D.1 prompts.

Results

Test A: Comprehension and Translation (from Dialect)

The results of test A.1 demonstrate ChatGPT's good level of proficiency in understanding the dialectal texts and translating them from dialect to Italian. As hypothesized, there seems to be a significant drop in performance with the high-temperature setting, while scores remain fairly consistent between the default and low temperature settings, pointing the default temperature setting probably as the most effective for the translation task. Upon analysing the variations across the three dialects, the performance appears to be superior for Roman and Sicilian dialects when evaluated using BLEU and TER scores, whereas the disparity is less evident in the COMET scores. However, the Roman dialect consistently scored higher across all metrics. This finding is consistent with expectations, considering Roman dialect's closer affinity to standard Italian compared to the other examined dialects.

In the following tables, I will report the scores for each attempt.[24]

	BLEU[25]			TER[26]			COMET[27]		
	DT	LT	HT	DT	LT	HT	DT	LT	HT
Milanese	0.54	0.51	0.37	0.65	0.63	0.50	0.88	0.87	0.84
Roman	0.72	0.67	0.40	0.78	0.77	0.56	0.93	0.92	0.85
Sicilian	0.65	0.71	0.49	0.75	0.77	0.60	0.85	0.85	0.81

Table 9. Scores for ChatGPT's translations from dialect to Italian at default (DT), low (LT), and high temperature (HT) settings.

[24] The values are rounded to two decimal places.

[25] The score has been smoothed using method 7 (https://www.nltk.org/api/nltk.translate.bleu_score.html), which returned the best results in this test.

[26] Values are normalized to a range of 0-1 to facilitate comparison with the other two scores.

[27] Only the system score is reported.

For test A.2, trying to provide a measurable assessment of the results, I manually assigned scores to ChatGPT's responses, ranging from 0 to 3 based on the following criteria:

0 = Misunderstanding of the text, resulting in an incorrect answer.
1 = Partial comprehension of the text, resulting in a partially correct answer.
2 = Good comprehension of the text, resulting in a correct answer.
3 = Optimal comprehension of the text, resulting in a complete and correct answer.

The scores in the following table represent the rating assigned for each of the five questions, with a maximum sum of 15 points. To mitigate potential biases arising from subjective judgement, I elaborated a checklist with concepts and keywords whose presence could mark the completeness of ChatGPT's answer.[28]

As evident from the table, the scores for after-translation answers are sometimes superior to those noted before translation. However, a more pronounced disparity is noted in terms of fluency and smoothness of writing style – a criterion not assessed in this context – which seemed to be better in post-translation answers. This outcome possibly arises from the fact that having access to the translated version permits a more focused approach to the subsequent task (answering the questions), facilitating more refined responses.

Generally, the test highlights ChatGPT's proficient understanding of dialectal texts. Surprisingly, the Roman dialect text garnered more superficial responses in the pre-translation stage, despite its closer resemblance to standard Italian. This phenomenon cannot easily be ascribed to a deficit in comprehension. In fact, as will also emerge in the findings from test C.1, it suggests the presence of underlying sociolinguistic biases influencing the perception and usage of the Roman dialect, potentially reflecting its perceived status among Italian speakers [50], [16]. Although this hypothesis demands further exploration, it is possible that the dialect's intrinsic prestige could have directed ChatGPT's responses to a less detailed analysis.[29]

	Milanese		Roman dialect		Sicilian	
	BT	AT	BT	AT	BT	AT
Question 1)	2	2	2	3	3	2
Question 2)	1	3	1	2	1	3
Question 3)	3	2	2	3	3	2

[28] Question 1: unknown seed; new specie of Baobab; sense of expectation. Question 2: being not ready; having just woken up; being metaphorically unkent. Question 3: long preparation; choosing colours and adjusting petals; connection with natural beauty of sun. Question 4: admiration despite the vanity; vulnerability; fascination. Question 5: taking care; breakfast; attention.

[29] This means that the prevalent informal usage of the Roman dialect, stemming from the perceived low sociolinguistic status of its speakers, might have influenced the output to be less refined and in-depth. Assuming that a more informal language usage can result in a diminished quality of task execution involving that language, even if debatable, it is intriguing to consider a potential correlation between the prestige of a language and ChatGPT's performance, particularly since the real-world materials it was trained on likely reflect the inherent biases in speakers' perceptions of languages. Thus, being trained predominantly on informal texts could have resulted in a superficial response in this comprehension task.

Question 4)	1	1	2	2	2	2
Question 5)	2	3	2	3	3	2
Total	9	11	7	13	12	11

Table 10. Scores achieved by ChatGPT in comprehension tests A.2, before translation (BT) and after translation (AT).

Test B: Discrimination and Analysis

Test B.1 underscored a general proficiency of ChatGPT in recognizing and distinguishing Italian dialects, even amongst less widespread subvarieties. In particular, the only misinterpretation occurred with the Calabrese dialect, which was initially confused with a Sicilian variant, due to their similarities. However, in the second call of the test,[30] the variety was correctly identified. Some uncertainty also arose when attempting to distinguish closely related subvarieties within the same group. For instance, when presented with a Bolognese text, ChatGPT correctly identified it as belonging to the Emilian group. Upon being asked to specify which subvariety, however, it initially misattributed it to Modenese in the first call, then correctly identifying it in the subsequent call.

Test B.2 aimed to further explore ChatGPT's ability to distinguish between subvarieties within the same dialectal group. As previously noted, the system displayed uncertainty when distinguishing between Bolognese and Modenese dialects. This can potentially be ascribed to the complex positioning of the Modenese dialect, which occupies a transitional zone between the Eastern and the Western group within the Emilian dialects, aligning with the former. When tasked with distinguishing it from the Bolognese dialect, ChatGPT exhibited considerable imprecision in its attribution. This trend persisted also when it was asked to identify the dialect of the Modenese text independently: while it correctly classified it as an Eastern variant (Parmesan) in the first call, it erroneously associated it with Ferrarese – a Western subvariety – in the second call. The performance markedly improved when distinguishing between the Milanese and Bergamo dialect, with the system accurately identifying them in the first call. In conclusion, it appears that the capability of correctly distinguishing between dialectal varieties depends, as expectable, on two factors: the similarity of the dialectal variety with other contiguous (Friulian better than Calabrese), and the level of spread and prestige that a dialect possesses (Milanese better than Bolognese). For task 2, which required highlighting the linguistic features that facilitate the differentiation, ChatGPT displayed null or very low analytical ability, being able only to identify the differences at a lexical level, and often misinterpreting morphological, phonological, and spelling traits.[31] This lack of expertise in analytical tasks is also confirmed by the following test.

Test B.3 was designed to evaluate ChatGPT's analytical skills by tasking it to identify a series of grammatical errors intentionally incorporated into texts in the three main dialect variants (Sicilian, Milanese, Roman dialect). Despite being prompted several times and guided toward the

[30] In instance of incorrect responses, negative feedback was provided, prompting a repetition of the task to seek different outcomes.

[31] For example, the system lists as phonetical features the use of symbols 'ʒ' to indicate a sound similar to 'j', or 'ó' to indicate an open 'o', clearly misunderstanding the difference between orthographic conventions and phonological traits.

error categories to focus on – such as grammatical dependencies – the system consistently failed to identify errors, frequently reporting either no issues or incorrect notations. These findings highlight ChatGPT's considerable limitations in performing analytical tasks on dialectal materials.

To determine a metric for the tests conducted in section B, I instituted a scoring system ranging from 0 to 1. A correct answer on the first attempt earned a score of 1, while a correct answer on the second attempt received a score of 0.5. In cases where both attempts were incorrect but contained some partially correct elements, a score of 0.25 was assigned.

B.1 (Dialect Recognition)	1st call	2nd call	Scores
Emilian (Bolognese)	Yes (but Modenese)	Yes (Bolognese)	1
Calabrian	No (Sicilian)	Yes	0.50
Friulian	Yes	-	1

Table 11. Results and corresponding scores for test B.1.

B.2 (Distinguishing subvarieties)	1st call	2nd call	Scores
1) Recognition: Bolognese – Modenese	No (Emilian – Brescian/Bergamo dialect)	No (Ferrara – Pavia dialect)	0.25
2) Analysis: Bolognese – Modenese	No	-	0
1) Recognition: Milanese – Bergamo	Yes	-	1
2) Analysis: Milanese – Bergamo	No (only lexical)	-	0.25

Table 12. Results and corresponding scores for test B.2.

B.3 (Errors Detection)	1st call	2nd call	Scores
Sicilian	No	No	0
Milanese	No	No	0
Roman	No	No	0

Table 13. Results and corresponding scores for test B.3.

Test C: Interaction and Translation (to Dialect)

In test C.1, the ability of ChatGPT to engage in a brief question-and-answer dialogue using a specific dialect showcased notable differences across each dialect, as well as identifiable trends in the system's behaviour across various linguistic contexts. The optimal performance was achieved when interacting with Milanese and Sicilian inputs; a marked enhancement was noted with Sicilian inputs when the specific linguistic domain of prompts was set. However, the usage of the dialects largely manifested in overlaying fundamentally Italian morphological and syntactical structures with phonetic elements typical of dialects, while the incorporation of

unique lexical items or syntactic markers remained minimal.[32] Essentially, the underlying framework was consistently Italian.

In the case of the Roman dialect, the system failed to maintain the dialect without a specific prompt instructing it to do so. Even with the specific prompt, it frequently reverted to standard Italian, demonstrating reduced consistency when the dialect shares greater similarity with the standard language. Another interesting observation pertains to the use of the Roman dialect: as noted earlier, the conversation generally adopted a less formal tone, enhanced by frequent use of idiomatic phrases which brought a distinctive colloquial flavour to the interaction.[33]

To evaluate ChatGPT's responses, I identified four parameters and for each of those I allocated scores ranging from 0 to 2, based on the following criteria:[34]

0 = Not or very low.
1 = Partially.
2 = Fully or very high.

[32] In the Milanese conversation, some characteristic features are consistently present, such as the dropping of final vowels other than [a] (e.g., 'temp'), the alteration of unstressed [i], especially in monosyllables (e.g., 'el', 'de'), and the simplification of double consonants to a single one (e.g., 'beli'). Morphosyntactic features include the use of the particle 'ghe' before the verb "to have" (e.g., 'te g'hai') and the plural determinative article 'i' in place of 'gli' before vowels. In the Sicilian conversation, notable phonetic features include the change of [o] to [u] and [e] to [i] both at end and in the middle of words. Morphologically, for instance, there's the usage of possessive pronouns (e.g., 'tò') and lexically of infinitive of verb 'essiri', as well as prepositions like 'cu' in place of 'con' or 'pi' for 'per', etc. Regarding the Roman dialect, we can observe phonetic traits such as the progressive assimilation of consonants from [nd] to [nn] or [rt] to [tt], the absence of diphthongization of Latin Ŏ (e.g., 'bono'), and the dropping of the final syllable in infinitive verb (e.g., 'esse', 'rilassà'). Morphological features of note include the use of the determinative masculine singular article 'er', and the plural 'li' in place of 'il' and 'gli'.

[33] For instance, consider sentences like «Se te piace magna', potresti annà a magna' una bella carbonara in un posto tipico romano», or «a Roma nun te se annoi mai!» which explicitly refer to the local Roman context, or the insertion of friendly allocution («amico mio») or proverbs («perché spesso, come se dice, "er denaro è er diavolo"». Such elements are noticeably absent in the other dialectal conversations.

[34] In this case, due to the complexity of the responses, I didn't utilize a check list for score attribution. Instead, I tried to assign an overall score based on the prevalence and frequency of dialectal traits. Naturally, this introduces a degree of scientific unreliability to the assigned scores, as they may vary with different evaluations. Nonetheless, this serves as suggestion for subsequent research, ideally whit the involvement of additional dialect experts.

Indicators	Sustained use of dialects	Incorporating dialect phonetics	Employing dialect morphology and syntax	Employing dialect lexicon	Total
Milanese[35]	2	2	1	0	5
Milanese (PSD)	2	2	1	0	5
Roman	0	0	0	0	0
Roman (PSD)	1	1	0	1	3
Sicilian	2	1	0	0	3
Sicilian (PSD)	2	2	0	1	5

Table 14. Scores achieved by ChatGPT in test C.1 (with and without PSD).

In test C.2, a counter-translation experiment was conducted, reverting from Italian back to dialect. The results were analysed using the same metrics as in test A.1, maintaining the default temperature setting, which had proven most effective. Notably, the outcomes here were significantly worse than in reverse translation, especially in the BLEU score, but slightly better when using my translations as reference. The Roman dialect yielded the best results, which – considering the outcomes of the previous tests – is possibly attributable to the close and extensive similarity between the linguistic features of the Roman dialect and standard Italian.

	BLEU		TER		COMET	
	MT	BT	MT	BT	MT	BT
Milanese	0.14	0.12	0.26	0.26	0.55	0.55
Roman	0.19	0.15	0.32	0.22	0.69	0.67
Sicilian	0.19	0.13	0.33	0.17	0.62	0.57

Table 15. Scores for ChatGPT translations from Italian to dialects using my translation (MT) or Bregoli's translation (BT) as reference.

Test D: Theoretical Background and Self-assessment

Regarding the skill tested in phase D, ChatGPT demonstrates substantial understanding of the distribution of the dialects and their subvarieties throughout Italy. For question 2, as already displayed by tests C, the system exhibits minimal ability to pinpoint the distinctive linguistic characteristics of each dialect, with frequent inaccuracies and mistakes, and often misunderstanding between phonetic elements and spelling. On the other side, ChatGPT demonstrates a commendable level of self-awareness regarding its knowledge base and the limitations in providing precise descriptions and analysis, for which human expertise is often

[35] It must be noted that ChatGPT's responses in the absence of a specific domain set aligned more with the Como's variety than the Milanese one, as evident by some orthographical conventions, like the letter *j* to indicate intervocalic *i* ([j]).

advised. Moreover, the AI showcases an optimal understanding of the areas needing improvement in its competency with Italian dialects, and of the methods by which such improvement could be achieved.

To provide a metric evaluation system, I manually assigned scores to ChatGPT's responses, ranging from 0 to 3, with a maximum score of 18 points, based on the following criteria:

0 = Incorrect answer.
1 = Partially correct answer, or incomplete answer.
2 = Correct answer and partially complete.
3 = Correct answer and fully complete.

In this case, a checklist was employed during the scoring process to reduce biases inherent in personal judgement.[36] However, score attribution could potentially vary with different evaluators. Despite this, instituting a metric evaluation was significant to obtain results suitable for comparison across the various tests conducted, and to offer a more clear and immediate sense of the system's performance.

	Milanese	Roman	Sicilian
Question 1)	2	3	3
Question 2)	1	0	1
Question 3)	2	2	1
Question 4)	2	2	2
Question 5)	3	2	3
Question 6)	3	3	3
Total	13	12	13

Table 16. Scores achieved by ChatGPT in test D.1.

[36] Question 1: identification of the specific geographic area, with awareness of the regional subvarieties. Question 2: accurate description of key features in phonetics, morphology, and syntax; recognition of a distinct lexicon and its influences. Question 3: recognition of established graphical conventions, their origins and application, as well as any similar attempts. Awareness of inconsistent graphic usage across different texts and authors. Question 4: detailed listing of the types of texts utilized as training sources. Question 5: accurate evaluation of system's capabilities and limitations concerning dialect proficiency. Question 6: comprehensive and accurate listing of potential sources for improvement.

Conclusion

In conclusion, the results obtained can be evaluated using two criteria: performance across different dialects, and performance based on the linguistic abilities assessed.[37] For each test, the scores were normalized to a range between 0 and 1, and the mean was then calculated. As observed in the resulting graph (Figure 1), ChatGPT's performance exhibits slight variation depending on the dialect, with the best outcomes observed for Sicilian, closely followed by Milanese, and lastly by the Roman. However, the patterns in skill performance were constant across dialects, suggesting that the results are more reflective of the system's inherent capabilities rather than related the specific dialect being tested.

Upon analysing the performance based on the skills tested (Figure 2), ChatGPT seems to possess a substantial ability to comprehend dialectal texts to a large extent.[38] It also showcases a commendable ability for recognizing various Italian dialects and often their subvarieties. However, when it comes to analysing the distinctive linguistic features that support this recognition ability, a significant shortfall in competency emerges, illustrating a gap in its analytical capacity.

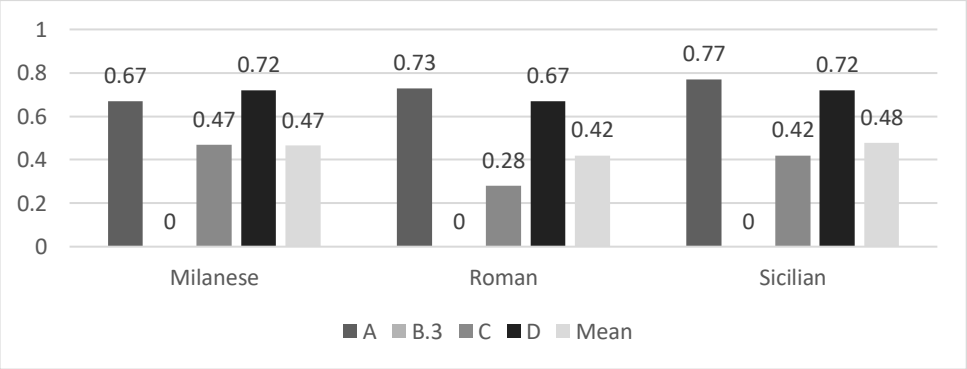

Figure 1. Results of the tests per dialect.

[37] In the case of test A.1, the results relative to the hight temperature setting was excluded. For the sum of the dialects score, tests B.1 and B.2 were excluded from the graph, given that they refer to other dialects.

[38] In fact, all the words in the texts have been understood, except for the Milanese 'scior', which means rich. In test C.1, it was replaced with 'rich' (a dialectal adaptation of Italian 'ricco') to keep the test going.

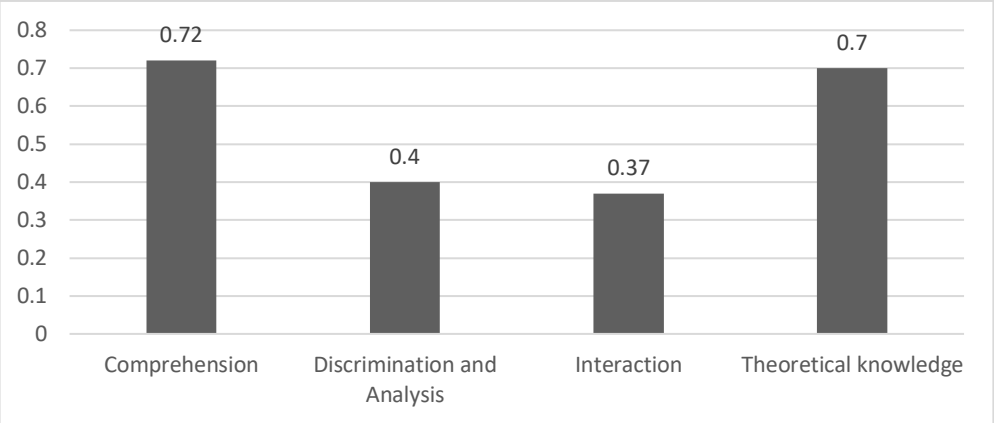

Figure 2. Results of the tests per skill.

To answer the initial question – "Does ChatGPT understand Italian dialects?" – it can be stated that ChatGPT proved a good understanding of these varieties, particularly in terms of passive linguistic competence. However, the system falls markedly short in active language engagement, underscoring the necessity of advancements in both interaction and production capacities. Potential improvements could be achieved through the expansion of the training dataset to include a broader array of dialectal materials, encompassing both written and spoken elements. Furthermore, training the system to undertake detailed analytical tasks could foster a deeper comprehension of different linguistic varieties, empowering it to identify and skilfully apply the distinctive characteristics of each dialect during communication tasks.

It is essential to highlight once more that the results presented in this study require validation through more systematic testing. Such testing should encompass a wider array of textual sources and dialectal varieties, and should involve multiple raters for the evaluative phase. Therefore, it is premature to make solid general assumptions from this single experiment. Nonetheless, my primary objective has been to outline a potential methodology for investigating ChatGPT linguistic competences, which may serve as model for humanities scholars seeking to conduct deeper and more extensive research in this domain.

References

[1] Alam, Md Mahfuz Ibn, Sina Ahmadi, and Antonios Anastasopoulos. 2023. "CODET: A Benchmark for Contrastive Dialectal Evaluation of Machine Translation". arXiv, Computing Research Repository. Last modified May 26, 2023. https://doi.org/10.48550/arXiv.2305.17267.

[2] Ascoli, Graziadio Isaia. 1882–5. "L'Italia dialettale". *Archivio glottologico italiano* 8: 98–128.

[3] Avolio, Francesco. 2009. *Lingue e dialetti d'Italia*. Roma: Carocci.

[4] Avolio, Francesco. 2010. "Laziali, Dialetti". Enciclopedia dell'Italiano, Istituto dell'Enciclopedia Italiana "Giovanni Treccani". https://www.treccani.it/enciclopedia/dialetti-laziali_(Enciclopedia-dell%27Italiano).

[5] Avolio, Francesco. 2011. "Siciliani, calabresi e salentini, Dialetti". Enciclopedia dell'Italiano, Istituto dell'Enciclopedia Italiana "Giovanni Treccani". https://www.treccani.it/enciclopedia/siciliani-calabresi-e-salentini-dialetti_%28Enciclopedia-dell%27Italiano%29.

[6] Bachman, Lyle F., and Adrian S. Palmer. 1996. *Language Testing in Practice. Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.

[7] Bachman, Lyle F., and Adrian S. Palmer. 2010. Language Assessment in Practice: Developing Language Assessments and Justifying Their Use in the Real World. Oxford: Oxford University Press.

[8] Bird, Steven, Ewan Klein, and Edward Loper. 2009. "Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit". O'Reilly Media, Inc. https://www.nltk.org/.

[9] Biswas, Som. 2023. "Evaluating Errors and Improving Performance of ChatGPT". *International Journal of Clinical and Medical Education Research* 2 (6): 182–8. https://dx.doi.org/10.33140/IJCMER.

[10] Bonfadini, Giovanni. 2010. "Lombardi, Dialetti". Enciclopedia dell'Italiano, Istituto dell'Enciclopedia Italiana "Giovanni Treccani". https://www.treccani.it/enciclopedia/dialetti-lombardi_(Enciclopedia-dell%27Italiano)/.

[11] Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, et al. 2020. "Language Models Are Few-Shot Learners". arXiv, Computing Research Repository. Last modified July 22, 2020. https://doi.org/10.48550/arXiv.2005.14165.

[12] Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, et al. 2023. "Sparks of Artificial General Intelligence: Early experiments with GPT-4". arXiv, Computing Research Repository. Last modified April 13, 2023. https://doi.org/10.48550/arXiv.2303.12712.

[13] Canale, Michael, and Merrill Swain. 1980. "Theoretical bases of communicative approaches to second language teaching and testing". *Applied Linguistics* 1 (1): 1–47. http://dx.doi.org/10.1093/applin/I.1.1.

[14] Cherubini, Francesco. 1840–3. *Vocabolario Milanese-Italiano*. Milano: Regia Stamperia.

[15] Circolo Filologico Milanese. 2018. *Vocabolario Milanese. Milanese-Italiano, Italiano-Milanese*, 3rd ed. Milano: Vallardi.

[16] D'Achille, Paolo, and Claudio Giovanardi. 2001. "Romanesco, neoromanesco o romanaccio? La lingua di Roma alle soglie del duemila". In *Dal Belli ar Cipolla. Conservazione e innovazione nel romanesco contemporaneo*. 13–28. Roma: Carocci.

[17] D'Achille, Paolo, and Claudio Giovanardi. 2016–8. *Vocabolario del Romanesco Contemporaneo*, 2 vols. Roma: Aracne.

[18] D'Achille, Paolo. 2002. "Il Lazio". In *I dialetti italiani. Storia, struttura, uso*, edited by Michele Cortelazzo et al., 515–58. Torino: UTET.

[19] de Saint-Exupéry, Antoine. (1943) 2014. *Il piccolo principe*. Traduzione di Nini Bompiani Bregoli. Milano: Bompiani.

[20] de Saint-Exupéry, Antoine. 2011. Er Principetto. Co le figure fatte co le mano de chi ha scritto er libbro. Ner parlà de noantri. Bologna: Massimiliano Piretti Editore.

[21] de Saint-Exupéry, Antoine. 2015. *El Princip Piscinin. Cont i acquarèi de l'autor. In lingua milanesa*, Bologna: Massimiliano Piretti Editore.

[22] de Saint-Exupéry, Antoine. 2023. U principinu. Cu l'acquarelli di l'autori. Traduzioni dû francisi 'nsicilianu di Mario Gallo. Neckarsteinach: Edition Tintenfass.

[23] Gao, Yuan, Ruili Wang and Feng Hou. 2023. "How to Design Translation Prompts for ChatGPT: An Empirical Study". arXiv, Computing Research Repository. Last modified April 21, 2023. https://doi.org/10.48550/arXiv.2304.02182.

[24] Haleem, Abid, Mohd Javaid, and Ravi Pratap Singh. 2022. "An Era of ChatGPT as a Significant Futuristic Support Tool: A Study on Features, Abilities, and Challenges". *BenchCouncil Transactions on Benchmarks, Standards and Evaluations* 2 (4), 100089. https://doi.org/10.1016/j.tbench.2023.100089.

[25] Han, Ridong, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. "Is Information Extraction Solved by ChatGPT? An Analysis of Performance, Evaluation Criteria, Robustness and Errors". arXiv, Computing Research Repository. Last modified May 23, 2023. https://doi.org/10.48550/arXiv.2305.14450.

[26] Hintze, Arend. 2023. "ChatGPT believes it is conscious". arXiv, Computing Research Repository. Last modified March 29, 2023. https://doi.org/10.48550/arXiv.2304.12898.

[27] Jiao, Wenxiang, Wenxuan Wang, Jen-tse Huang, Xing Wang and Zhaopeng Tu. 2023. "Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine". arXiv, Computing Research Repository. Last modified May 19, 2023. https://doi.org/10.48550/arXiv.2301.08745.

[28] Kane, Micheal T. 2006. "Validation". In *Educational Measurement*, 4th ed., edited by Robert L. Brennan, 17–64. Washington, DC: Rowman & Littlefield.

[29] Kane, Micheal T. 2013. "Validating the interpretations and uses of test scores". *Journal of Educational Measurement* 50 (1): 1–73. https://doi.org/10.1111/jedm.12000.

[30] Kocoń, Jan, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydlo, Joanna Baran and Julita Bielaniewicz. 2023. "ChatGPT: Jack of all trades, master of none". *Information fusion* 99 (November). https://doi.org/10.1016/j.inffus.2023.101861.

[31] Koubaa, Anis, Wadii Boulila, Lahouari Ghouti, Ayyub Alzahem, and Shahid Latif. 2023. "Exploring ChatGPT Capabilities and Limitations: A Critical Review of the NLP Game Changer". *Preprints* 2023. Last modified March 23, 2023. https://doi.org/10.20944/preprints202303.0438.v1.

[32] Lai, Viet Dac, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. "ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning". arXiv, Computing Research Repository. Last modified April 12, 2023. https://doi.org/10.48550/arXiv.2304.05613.

[33] Li, Bo, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023. "Evaluating ChatGPT's Information Extraction Capabilities: An Assessment of Performance, Explainability, Calibration, and Faithfulness". arXiv, Computing Research Repository. Last modified April 23, 2023. https://doi.org/10.48550/arXiv.2304.11633.

[34] Li, Linhan, Huaping Zhang, Chunjin Li, Haowen You, and Wenyao Cui. 2023. "Evaluation on ChatGPT for Chinese Language Understanding". *Data Intelligence* 5 (4). https://doi.org/10.1162/dint_a_00232.

[35] Loporcaro, Michele. 2013. *Profilo linguistico dei dialetti italiani*. Roma / Bari: Laterza.

[36] Merlo, Clemente. 1960–1. "I dialetti lombardi". *L'Italia dialettale* 24: 1–12.

[37] Messick, Samuel. 1989. "Validity". In *Educational Measurement*, 3rd ed., edited by Robert L. Linn 13–100. Washington, DC: Macmillan / American Council on Education.

[38] OpenAI. 2023. "GPT-4 Technical Report". arXiv, Computing Research Repository. Last modified March 27, 2023. https://doi.org/10.48550/arXiv.2303.08774.

[39] Papineni, Kishore, Salim Roukos, Todd Ward, Wei-Jing Zhu. 2002. "BLEU: a method for automatic evaluation of machine translation". In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. 311–8. Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135.

[40] Pareschi, Remo. 2023. "Abductive reasoning with the GPT-4 language model: Case studies from criminal investigation, medical practice, scientific research". *Sistemi intelligenti* 25 (2): 435–444. https://doi.org/10.1422/108139.

[41] Pellegrini, Giovanni Battista. 1975. "I cinque sistemi dell'italo-romanzo". In *Saggi di linguistica italiana. Storia, struttura, società*. 55–87. Torino: Boringhieri.

[42] Peng, Keqin, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. "Towards Making the Most of ChatGPT for Machine Translation". arXiv, Computing Research Repository. Last modified March 24, 2023. https://doi.org/10.48550/arXiv.2303.13780.

[43] Piccitto, Giorgio, ed. 1977–2002. *Vocabolario Siciliano*, 5 vols. Palermo / Catania: Centro di Studi Linguistici e Filologici Siciliani.

[44] Ray, Partha Pratim. 2023. "ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope". *Internet of Things and Cyber-Physical Systems* 3: 121–54. https://doi.org/10.1016/j.iotcps.2023.04.003.

[45] Rei, Ricardo, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. "COMET: A Neural Framework for MT Evaluation". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2685–702. Stroudsburg, PA: Association for Computational Linguistics. https://aclanthology.org/2020.emnlp-main.0.pdf.

[46] Roumeliotis, Konstantinos I., Nikolaos D. Tselikas. 2023. "ChatGPT and Open-AI Models: A Preliminary Review". *Future Internet* 15 (6), 192. https://doi.org/10.3390/fi15060192.

[47] Ruffino, Giovanni. 1984. "Isoglosse siciliane". In Tre millenni di storia linguistica della Sicilia. Atti del Convegno della Società italiana di glottologia (Palermo, 25-27 marzo 1983), edited by Adriana Quattordio Moreschini, 161–224. Pisa: Giardini.

[48] Sanga, Glauco. 1984. *Dialettologia lombarda. Lingue e culture popolari*. Pavia: Università di Pavia, Dipartimento di Scienza della Letteratura.

[49] Sanga, Glauco. 1999. "Il dialetto di Milano". *Rivista italiana di dialettologia* 23: 137–64.

[50] Serianni, Luca. 1987. "Lingua e dialetto nella Roma del Belli". *Studi Linguistici Italiani* XIII: 204–21.

[51] Snover, Matthew, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. "A Study of Translation Edit Rate with Targeted Human Annotation". In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*. 223–31. Cambridge, Massachusetts, USA: Association for Machine Translation in the Americas. https://aclanthology.org/2006.amta-paper s.25.

[52] Sohail, Shahab Saquib, Faiza Farhat, Yassine Himeur, Mohammad Nadeem, Dag Øivind Madsen, Yashbir Singh, Shadi Atalla, and Wathiq Mansoor. 2023. "Decoding ChatGPT: A Taxonomy of Existing Research, Current Challenges, and Possible Future Directions". arXiv, Computing Research Repository. Last modified August 25, 2023. https://doi.org/10.48550/arXiv.2307.14107.

[53] Stahl, Bernd C., and Damian Eke. 2024. "The ethics of ChatGPT – Exploring the ethical issues of an emerging technology". *International Journal of Information Management* 74, 102700. https://doi.org/10.1016/j.ijinfomgt.2023.102700.

[54] Trifone, Pietro. 1992. *Roma e il Lazio*. Torino: UTET.

[55] Trovato, Salvatore C. 2002. "Sicilia". In *I dialetti italiani. Storia, struttura, uso*, edited by Michele Cortelazzo et al., 834–97. Torino: UTET.

[56] Widdowson, Henry .G. 1983. *Learning Purpose and Language Use*. Oxford: Oxford University Press,

[57] Zeng, Fankun. 2023. "Evaluating the Problem Solving Abilities of ChatGPT". Master Thesis, Master of Science, McKelvey School of Engineering, Washington University in Saint Luis, Spring 5-2023. McKelvey School of Engineering Dissertation & Thesis (849). https://doi.org/10.7936/7vz0-dr08.

[58] Zhou, Jianlong, Heimo Müller, Andreas Holzinger, and Fang Chen. 2023. "Ethical ChatGPT: Concerns, Challenges, and Commandments". arXiv, Computing Research Repository. Last modified May 18, 2023. https://doi.org/10.48550/arXiv.2305.10646.