Benchmarking AI acceptability and grammaticality in German: A study of ChatGPT and human judgments

Nicholas Catasso (Bergische Universität Wuppertal)

catasso(at)uni-wuppertal.de

Abstract

The rapid development of large language models has opened new avenues for linguistic research, including areas traditionally reliant on native-speaker intuitions. One such domain is grammaticality and acceptability judgment, where speakers assess whether sentences are structurally well-formed and contextually appropriate. This study investigates the extent to which ChatGPT-4 can approximate human judgments in German, focusing on a diverse range of grammatical and usage-related phenomena. A carefully designed set of test items was presented to both the model and native speakers, allowing for a direct comparison. The results show a high degree of alignment in many cases, but also reveal systematic divergences, particularly in contexts involving gradience, sociolinguistic markedness or context-dependent acceptability. These findings demonstrate both the analytical potential and the current limitations of large language models in linguistic research, and contribute to ongoing discussions about their ability to approximate native speaker competence.

Keywords

LLMs, ChatGPT, grammaticality, acceptability, German

1 Introduction

In recent years, Artificial Intelligence (AI) has achieved remarkable advances across a wide range of disciplines. Within linguistics, the integration of AI applications – most prominently large language models (LLMs) – has opened new avenues for the analysis of natural language (see, among many others, Austin et al. 2021; Bang et al. 2023; Albrecht 2023; Liu et al. 2023; Tikhonova & Raitskaya 2023; Yu et al. 2024, Meier 2024, Masood & Khan 2025, De Cesare et al. (eds) 2025). From morphosyntax and semantics to variationist studies, these technologies provide linguists with unprecedented opportunities to investigate complex patterns of language structure and use. Tasks such as text generation, translation and corpus annotation can now be automated on a scale and with a degree of sophistication that would have been inconceivable only a decade ago. Prior research has examined various aspects of older models, including their capacity for generating coherent text (e.g., Radford et al. 2019; Brown et al. 2020, Li 2022) and understanding natural-language syntax (e.g., Zhou et al., 2023). Additionally, more recent studies have investigated their performance on specific linguistic tasks such as well-formedness judgments (e.g. Peters et al. 2018; Qiu et al. 2024; Hu et al. 2024; Leivada et al. 2024; Johnsen 2025; Ide et al. 2025).

These developments, however, are accompanied by substantial challenges. A central concern is the availability of appropriate training data that adequately

AILing

Catasso, Nicholas. 2026.
Benchmarking AI acceptability and grammaticality.
Special Issue: *Natural Language and AI*. Vol.3 No.1
DOI: 10.62408/ai-ling.v3i1.35.

AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses ISSN: 2943-0070

reflect the diversity of linguistic structures and phenomena. In practice, datasets are often incomplete, unevenly distributed or error-prone, which raises questions about the reliability of model outputs. Moreover, since LLMs are trained on existing language data, they are susceptible to reproducing the biases encoded in those data, thereby introducing systematic distortions into linguistic analyses. The issue of linguistic variation poses an additional difficulty: for models to serve as reliable tools for linguistic research, they must be sensitive to variation across registers, dialects and contexts. Yet in many cases, their training encourages a preference for standardized forms or frequent patterns, limiting their capacity to capture the full richness of natural language.

Against this backdrop, the present study investigates grammaticality and acceptability judgments in German. While LLMs such as ChatGPT have increasingly been evaluated with respect to their alignment with human judgments in English (cf., e.g. Marvin & Linzen 2018; Qiu et al. 2024; Bavaresco et al. 2025), there are, to the best of my knowledge, no comparable studies for German. This absence is noteworthy, given that grammaticality and acceptability judgments constitute a cornerstone of linguistic theory and empirical research. By addressing this gap, the study contributes to ongoing debates about the extent to which LLMs can approximate human linguistic competence in languages other than English.

The structure of the paper is as follows: Section 2 introduces the theoretical background and clarifies the core concepts of grammaticality and acceptability. Section 3 presents a pilot study designed to evaluate ChatGPT running on the original GPT-4 model (March 2023 release; Open AI 2023, henceforth: ChatGPT-4) ability to produce grammaticality and acceptability judgments for German, in direct comparison with a human control group. Section 4 reports and analyzes the results of this study. In Section 5, selected edge cases and particularly illustrative examples are discussed in greater depth. Finally, Section 6 summarizes the key findings and reflects on their broader implications.

2 Grammaticality and acceptability

The distinction between grammaticality and acceptability has played a central role in linguistic theory and empirical research for decades. While the two concepts are closely related, they are not interchangeable. Grammaticality is typically understood as a property of linguistic form, whereas acceptability concerns speakers' judgments about the naturalness or appropriateness of utterances in context. In what follows, we briefly outline these two notions and highlight their theoretical and methodological implications.

2.1 Grammaticality

In the generative tradition, the notion of grammaticality designates the degree to which a string of words conforms to the formally specified rules and principles of

a native speaker's internalized grammar (Chomsky 1965, 1986). A sentence is judged grammatical if it can be generated by the grammar, and ungrammatical if it falls outside its derivational space. Traditionally, grammaticality has been modeled as a binary feature, with sentences assigned either [+grammatical] or [-grammatical] values (Levelt et al. 1977; Derwing 1979; Cohen 1981; Gross 2021).

In German, the contrast in (1) exemplifies this view. The finite verb in a declarative root clause is subject to the Verb-Second (V2) constraint, which requires it to appear in the second position of the clause, immediately following exactly one constituent in the so-called *Vorfeld* ('prefield'):

(1) a. Heute hat Maria ein Fahrrad gekauft.
today AUX.IND.PRS.3SG Maria INDEF.ACC.SG.M bicycle PTCP-have-PTCP

b. *Maria heute ein Fahrrad gekauft hat.
Maria today INDEF.ACC.SG.M bicycle PTCP-have-PTCP AUX.IND.PRS.3SG
'Maria bought a bicycle today.'

Here, (1a) is grammatical, whereas (1b) is ungrammatical (*) precisely because it violates this linearization requirement.

However, empirical research has shown that judgments of grammaticality are not always clear-cut. Even when sentences diverge from canonical rules, speakers' intuitions may vary, and some structures may elicit gradient or context-dependent evaluations. As Featherston (2007), Schütze (2016) and Sprouse (2018) argue, what is traditionally labeled "grammaticality" interacts with performance factors, processing constraints and inter-individual variation. In some approaches, it is even argued that grammaticality is "not directly accessible to observation or measurement" (Lau et al. 2016: 3). Thus, while grammaticality is theoretically conceived as a categorical notion, its empirical assessment frequently exhibits gradience[1] and variation in speakers' judgments.

[1] In the present discussion, the term "gradience" refers to the non-binary nature of acceptability judgments, whereby speakers evaluate linguistic structures along a continuum ranging from fully acceptable to unacceptable. Such gradient patterns may reflect the influence of processing, contextual, and inter-individual factors, even if grammatical well-formedness is theoretically conceived as categorical.

2.2 Acceptability

Acceptability, by contrast, is best understood as the epistemic correlate of native speakers' intuitions regarding the naturalness, felicity or contextual appropriateness of an utterance. Whereas grammaticality constitutes an abstract, competence-level property defined over the derivational possibilities afforded by an internalized grammar, acceptability reflects interface-level evaluations of linguistic forms as they are parsed, interpreted and embedded in discourse. Acceptability judgments thus arise from the interaction of morphosyntactic well-formedness with pragmatic felicity conditions, discourse coherence constraints, stylistic conventions and socio-cultural norms (Grice 1975; van Dijk 1977; Sperber & Wilson 1998).

The correlation between grammaticality and acceptability is non-isomorphic. Utterances may be syntactically well-formed yet pragmatically infelicitous – for example, when they violate conversational maxims, clash with world knowledge or conflict with register-specific norms. Conversely, strings that deviate from canonical grammatical rules may nevertheless be judged acceptable in situated discourse, insofar as they are licensed by communicative intent, processing efficiency or usage-based conventionalization.

This asymmetry is illustrated in (2). The well-known sentence in (2a), though fully grammatical, is semantically anomalous and thus judged unacceptable (marked with #). In contrast, (2b) features a morphosyntactic violation – specifically, the use of the accusative case (*einen Mann*) where the nominative (*ein Mann*) is expected following the copula *ist* ('is') – yet is readily accepted by native speakers in spontaneous speech (the example is drawn from a written corpus of informal online communication) and may be associated with high acceptability in such contexts. One plausible explanation for this pattern is that speakers do not immediately interpret the sentence as a copular construction. Instead, *Was ich will ist* ... is initially processed as a main clause (*Ich will…* 'I want…'), in which *einen Mann* serves as the canonical object of the verb *will*. This real-time parsing strategy leads to an interpretation effectively masking the case conflict:

(2) a. [+grammatical, -acceptable]

#Farblose	grüne	Ideen
colorless-NOM.PL.S	green-NOM.PL.S	idea-NOM.PL
schlafen	wütend.	
sleep-IND.PRS.3PL	furiously	

'Colorless green ideas sleep furiously.'
(Chomsky 1957: 15, transl. as in the source)

b. [-grammatical, +acceptable]

Was	ich	will	ist	
what-ACC	I-NOM	want-IND.PRS.1SG	be-IND.PRS.3SG	
einen	Mann	an	meiner	Seite …
INDEF.ACC.SG.M	man	at	my-DAT.SG.F	side

> ‘What I want is a man by my side …’
> (DWDS, Webkorpus, ~2017-07-09, punctuation unmodified, my transl.)[2]

From a theoretical standpoint, acceptability is inherently gradient, context-sensitive, and probabilistic. While both grammaticality and acceptability can exhibit gradience and variation in empirical judgment data, they differ in their underlying sources: grammaticality reflects properties of the internalized competence grammar, whereas acceptability emerges from the interaction of that grammar with performance systems and contextual constraints. Acceptability judgments therefore serve as a crucial empirical window into the interface between syntactic knowledge and pragmatic, cognitive, and usage-based factors.

3 The study

This section presents a study designed to evaluate how closely ChatGPT-4 approximates native speaker judgments of grammaticality and acceptability in German. By directly comparing model-generated evaluations with those of human participants, the study assesses both the potential and the limitations of LLMs in tasks that rely on linguistic intuition. The chapter outlines the guiding research questions and hypotheses (Section 3.1), describes the materials and participant groups (Section 3.2) and details the construction and categorization of the stimulus items (Section 3.3).

3.1 Research questions and hypotheses

The research questions reflect two central aims: to measure the extent to which ChatGPT-4's grammaticality and acceptability judgments align with those of native speakers (Q1) and to identify possible factors that may account for divergences between them (Q2). These questions are situated within the broader debate about whether LLMs can serve as reliable proxies for human linguistic competence:

Q1: How accurately does ChatGPT-4 generate grammaticality judgments compared to human participants? What conclusions can be drawn from this comparison?
Q2: What factors influence the validity of grammaticality and acceptability judgments generated by ChatGPT-4 compared to those made by humans?

[2] To be sure, the status of sentences that are ungrammatical yet perceived as acceptable has been the subject of sustained debate. Some scholars view such cases as by-products of processing constraints, while others argue that they reflect grammatical structures whose acceptability is modulated by contextual or performance-related factors. For a comprehensive overview, see, e.g., Gibson & Thomas (1999); Wellwood et al. (2018); Leivada and Westergaard (2020).

To address Q1, the study tests the following hypotheses:

H_0: There are no significant differences between the grammaticality/acceptability judgments of ChatGPT-4 and those of native German speakers.
H_1: There are significant differences between the grammaticality/acceptability judgments of ChatGPT-4 and those of native German speakers.

As to Q2, this question cannot be addressed through quantitative analysis on the basis of the present dataset and experimental design, i.e., the available data do not permit a statistically robust modeling of the factors potentially underlying divergences between human and model-based judgments. Instead, Q2 is approached through a qualitative, interpretative analysis that situates observed discrepancies within a broader theoretical and methodological framework.
This qualitative perspective allows the study to move beyond mere performance comparison and to engage with the deeper issue of what it means for an LLM to "judge" linguistic well-formedness, thereby contributing to the broader debate on the status of LLM outputs as evidence for linguistic competence.

3.2 Methods

The stimulus set comprised 60 contextualized German utterances, evenly distributed across five experimental conditions (12 items per group). Each group was designed to test a distinct grammatical or sociolinguistic dimension, including canonical and non-canonical syntax, dialectal and register variation and contested or marginal constructions (see Section 3.3 for full categorization). The entire study was conducted in German, including all materials and instructions. All sentences were presented in minimal contextual framing, sufficient to disambiguate their intended reading while avoiding pragmatic overload.[3]

Participants were asked to evaluate each utterance according to their spontaneous linguistic intuition concerning its perceived naturalness and correctness in everyday German. The evaluation task was modeled on System 1 processing (Kahneman 2011), aiming to elicit rapid, intuition-driven responses rather than reflective or rule-based judgments. Sentence ratings were recorded on a five-point Likert scale ranging from 1 (= 'completely unnatural') to 5 ('completely natural') (Figure 1):

[3] The complete experimental materials are not reproduced in the article due to space limitations. For purposes of transparency and methodological documentation, all materials are deposited in an open-access repository. This includes (i) the initial prompt used to elicit participation from ChatGPT-4, (ii) the instructions presented to both human participants and the AI condition, and (iii) the full set of 60 contextualized German stimuli employed in the acceptability judgment task. The dataset is available via Zenodo at https://zenodo.org/records/18247009 (DOI: 10.5281/zenodo.18247008).

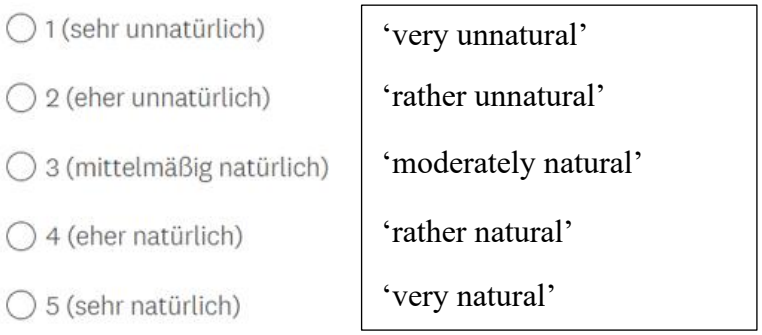

Figure 1. Five-point Likert scale used for ratings, ranging from 1 ('very unnatural') to 5 ('very natural').

The instructions explicitly emphasized that participants should rely on their internal sense of linguistic appropriateness rather than appeal to formal grammatical rules, normative prescriptions or orthographic conventions, and that sentences were to be evaluated with respect to the context provided.

The same set of stimuli was submitted to two evaluation conditions, designed to allow a direct comparison between model-generated judgments and human intuitions. Study 1 (S1) involved 65 monolingual native speakers of German,[4] all born and residing in the federal state of North Rhine-Westphalia (western Germany) at the time of the data collection. This population was selected to control for potential confounds arising from dialectal heterogeneity and multilingual influence. While features of Western Middle German are present in the regional vernacular, North Rhine-Westphalia is not characterized by dialect dominance and variation to the same degree as other German-speaking regions, particularly among the demographic sampled in this study (see below). Furthermore, restricting participation to monolingual speakers reduced the likelihood of transfer effects from other languages in the judgment task. Crucially, the regionally compact cohort provided a shared linguistic background, yielding a maximally uniform baseline for intuitions about grammaticality and linguistic variation. The sample included 43 women, 20 men and 2 individuals identifying as diverse, with a mean age of 22.3 years. Participants were recruited via academic mailing lists, university networks and local contacts. Involvement in the study was voluntary and uncompensated. The data were collected anonymously through an online questionnaire administered via the Survio platform (https://www.survio.com/de/). In Study 2 (S2), the full set of 60 utterances was

[4] The term "monolingual" is used in a strict sense here: all participants reported German as their only L1 and did not acquire any additional native language(s) during early childhood. This restriction was imposed to maximize experimental control and to minimize potential confounds related to early bilingualism or multilingual language exposure.

input into ChatGPT-4 in a single prompt.[5] To avoid contamination effects, no item-specific feedback, clarification prompts or iterative refinements were provided. All responses were generated in a single session, with the model directly prompted to provide ratings on the same five-point Likert scale used in the human evaluation task. These outputs were treated as scalar judgments for purposes of comparison. While ChatGPT-4 does not possess intuitive linguistic knowledge in the human sense – that is, it does not have access to the kind of embodied, experiential competence often associated with human speakers[6] – its outputs can nonetheless approximate human judgments through statistical learning on large-scale language data. Evaluating its performance on the same task as human speakers allows for a controlled investigation into the extent to which its output converges with native-speaker intuitions, and where systematic divergences emerge.

All participants in S1, like ChatGPT-4 in S2, received identical instructions designed to elicit spontaneous, intuition-based judgments rather than deliberate linguistic analysis. The sentences were presented in randomized order and participants were encouraged to respond spontaneously, without time constraints, to approximate natural language processing conditions. The use of a uniform stimulus set, a consistent five-point scalar rating format, and parallel task framing across both studies served to maximize comparability between human and model responses while minimizing extraneous methodological variability. This experimental symmetry not only enables a direct comparison of aggregate acceptability ratings, but also supports a fine-grained analysis of how specific grammatical and sociolinguistic factors modulate alignment between native speaker intuition and LLM output.

[5] As an anonymous reviewer points out, one potential methodological concern is whether the specific way in which the stimuli were submitted to ChatGPT-4 in Study 2 (namely, presenting the full set of items within a single conversational prompt) may have influenced the model's judgments. To assess this possibility, an exploratory follow-up test was conducted in which a subset of the stimuli (25 items) was submitted to ChatGPT-4 in a separate conversation, using the same instructions but a different session and device, and at approximately the same time as the original run. The outputs obtained in this alternative setup were then compared with those from the original configuration. This comparison did not reveal any systematic differences in the model's grammaticality or acceptability judgments. Within the limits of this exploratory check, the results therefore do not appear to be driven by the particular conversational configuration employed in Study 2.

[6] An anonymous reviewer notes that this formulation might invite different interpretations and therefore warrants clarification. In the present context, this expression is not meant to deny that ChatGPT-4 has been trained on large amounts of linguistic data. Rather, it refers to a distinction commonly drawn in linguistics and cognitive science between linguistic knowledge grounded in embodied, situated, and socially mediated experience and knowledge derived from statistical regularities in textual input. While ChatGPT-4 has extensive exposure to language through pre-training, it does not possess sensorimotor experience, situational awareness, or participation in communicative practice. The claim in the main text thus concerns the absence of embodied and situated linguistic experience, not the absence of large-scale exposure to linguistic data.

3.3 Stimuli

The stimulus set was organized into five distinct groups, each targeting a specific type of linguistic variation. This design was intended to probe not only the AI model's and participants' sensitivity to grammatical well-formedness, but also their responsiveness to sociolinguistic and contextual variation. The five groups were as follows:

Table 1: Stimuli in S1 and S2.

GROUP	LABEL	DEFINITION	DESCRIPTION
G1	grey-zone items	grammatical constructions with variable acceptability judgments among native speakers.	productively used forms that often elicit inconsistent intuitions due to norm conflict, optionality, or stylistic markedness.
G2	grammatical items	sentences that are uncontroversially grammatical and pragmatically neutral.	canonical structures that conform to prescriptive norms and are fully acceptable across dialects, registers, and discourse contexts.
G3	ungrammatical items	sentences exhibiting categorical violations of morphosyntactic constraints.	includes errors in word order, agreement, case assignment, and other core syntactic domains of Standard German.
G4	diatopically marked items	sentences featuring regionally specific lexical or grammatical forms.	grammatical in regional varieties but perceived as marked or non-standard by speakers from other dialect areas.
G5	diastratically / diaphasically marked items	sentences containing socially or stylistically marked features.	characteristic of particular social groups or informal registers (e.g., youth language, colloquial speech), and potentially inappropriate in formal contexts.

G1 comprised constructions that are in principle grammatical and productively used in contemporary German, but whose acceptability is subject to interspeaker variability. Such structures (so-called *Zweifelsfälle* in German) occupy a gradient zone between prescriptive norms and actual usage, which makes them ideal for probing the model's alignment with speaker intuition in contexts of norm conflict, register tension, or optionality. Examples include constructions involving optional complementizers, syntactic extraposition, or variable word order patterns that are not outright ungrammatical but often perceived as marked or marginal.

G2 consisted of uncontroversially grammatical and pragmatically unmarked sentences that are fully acceptable across dialects, registers and discourse contexts. These items served as a baseline condition, providing a reference point for assessing both the reliability of participant ratings and the internal consistency of the model's judgments.

The items in G3 exhibited clear violations of core morphosyntactic constraints of German. These included, for instance, case mismatches, incorrect agreement patterns or violations of verb placement rules. The inclusion of this group was essential for evaluating the model's ability to detect categorical

grammatical violations and for contrasting categorical unacceptability with more gradient forms of deviance represented in other groups.

The G4 stimuli featured lexical and syntactic constructions characteristic of specific German dialect regions. While typically grammatical within their local varieties, these forms are often perceived by speakers from other regions as marked, non-standard or unacceptable. Including such items made it possible to examine how both human participants and the model respond to regionally restricted features, and whether their acceptability judgments reflect dialect familiarity or alignment with perceived standard norms.

Finally, G5 comprised utterances marked by features associated with particular social groups, communicative registers or stylistic domains – for example, youth language, colloquial spoken German or stylistically high or low varieties. These stimuli are all grammatical, but subject to variation in perceived appropriateness depending on discourse context. As such, they are not evaluated in terms of morphosyntactic well-formedness, but rather in terms of sociopragmatic fit and register alignment. The inclusion of this group was intended to isolate the role of contextual appropriateness in acceptability judgments. To that end, the group was internally balanced: six items were presented in discourse contexts that were socially and pragmatically appropriate for the register and style of the utterance (e.g., informal language in an informal setting), whereas the remaining six were embedded in contexts that were deliberately incongruent with these features (e.g., informal language in a highly formal setting). This design created a controlled contrast between contextually appropriate and contextually inappropriate uses of otherwise grammatical forms.

As noted above, all 60 items across the five groups were contextually framed – i.e., each was accompanied by a concrete and pragmatically coherent discourse scenario designed to support interpretation without introducing extraneous bias. These contextualizations grounded each stimulus in a plausible communicative setting, allowing both human participants and the model to evaluate the utterances with respect to natural usage conditions. The internal structure of the stimulus set was carefully balanced to ensure that no single dimension of variation dominated the task, thereby enabling a differentiated assessment of both grammatical and sociolinguistic dimensions of acceptability.

The following section provides illustrative examples from each stimulus group, clarifying the linguistic phenomena and variation types targeted in the design.

3.3.1 G1: Grey-zone items

Two variants of the same utterance, both included as distinct test items in the study and presented in randomized order, are illustrated in (3a) and (3b). Both were presented with the same contextual prompt, reproduced below in its original form and accompanied by an English translation. The example is also glossed for clarity:

(3) Sie rufen Hans an und hören laute Geräusche im Hintergrund. Sie fragen ihn: „Was sind diese Geräusche?“. Er antwortet:
‘You call Hans and hear loud noises in the background. You ask him: "What are these noises?" He replies:’

a. Ich staubsauge gerade den Teppich.
I-NOM dust-suck-IND.PRS.1SG right-now DEF.ACC.SG.M carpet

b. Ich sauge gerade den Teppich staub.
I-NOM suck-IND.PRS.1SG right-now DEF.ACC.SG.M carpet dust
‘I’m just vacuuming the carpet.’

The linguistic uncertainty in this case lies in the variation among native speakers regarding the morphosyntactic status of the verb *staubsaugen* (‘to vacuum’). While both variants are attested in actual usage, speakers differ systematically in their intuitions and production preferences. Some speakers treat *staubsaugen* as a morphologically simple, non-separable verb, analogous to a prefixed form in which *staub-* (‘dust’) functions either as a bound morpheme or even as a non-decomposable element of the verbal root (cf. (3a)). Others analyze the verb as a separable compound, decomposing it into the verbal stem *saug-* (‘suck’) and the particle-like element *staub* (‘dust’), with the latter surfacing in clause-final position (as in (3b)).[7] Still others reject both surface variants entirely and use the verb only in its clause-final infinitival form (e.g., *Heute muss ich staubsaugen*, lit. ‘today must

[7] Cf. other canonical German verbs such as *verstehen* ‘understand’ (i)-(ii) and *aufstehen* ‘get up’ (iii)–(iv), which exemplify these two morphosyntactic patterns (in (ii), the glosses are comparatively constructed to reflect the parallel status of the two components of the verb *understand* in English):

(i) Ich verstehe dich nicht.
I-NOM understand-IND.PRS.1SG you-ACC.SG NEG
‘I don’t understand you.‘

(ii) *Ich stehe dich nicht ver.
I-NOM stand-IND.PRS.1SG you-ACC.SG NEG ‘under’
(intended:) ‘I don’t understand you.’

(iii) Ich stehe jeden Tag um 7 Uhr auf.
I-NOM get-up-IND.PRS.1SG every-ACC.SG.M day at 7 hour V.PRT
‘I get up every day at 7 o’ clock.’

(iv) *Ich aufstehe jeden Tag um 7 Uhr.
I-NOM V.PRT-get-up-IND.PRS.1SG every-ACC.SG.M day at 7 hour
(intended:) ‘I get up every day at 7 o’ clock.’

Examples (i)-(ii) show that *verstehen* contains the inseparable prefix *ver-*, which is morphologically bound to the verb stem and cannot be detached. It obligatorily surfaces in clause-second position as part of the finite verb form. In contrast, (iii)-(iv) demonstrate the behavior of the separable verb *aufstehen*, in which the particle *auf-* is split off in main clauses and surfaces in clause-final position, as dictated by the V2 constraint in German.

I vacuum' = 'I have to vacuum today'), without producing finite conjugations of either type in spontaneous speech. For this reason, such verbal word formations are often referred to as "non-V2 verbs" (Freywald & Simon 2007; Forche 2020), "movement-resistant verbs" (Fortmann 2007, 2015) or "prefield-phobic verbs" (Sternefeld 2008; Ahlers 2010; Meinunger 2022). This type of variation exemplifies a productive grey-area phenomenon in contemporary German verb morphology, where competing morphological parses coexist within the speech community and give rise to divergent acceptability judgments (Vikner 2005; Freywald & Simon 2007). This pattern is likely linked to the relatively recent lexicalization of *staubsaugen* as a result of univerbation from the compositional phrase *Staub saugen* (lit. 'suck dust'), in which *Staub* originally functioned as the preverbal direct object – consistent with canonical SOV word order in German infinitival constructions.
The Google Ngram (Michel et al. 2011a, 2011b) data in Figure 2 illustrate that *staubsaugen* only began to appear with notable frequency in written German during the second half of the 20th century. Its comparatively recent emergence and ongoing grammaticalization may account for its incomplete syntactic stabilization within the language system and thus for the persistent variation observed among speakers:

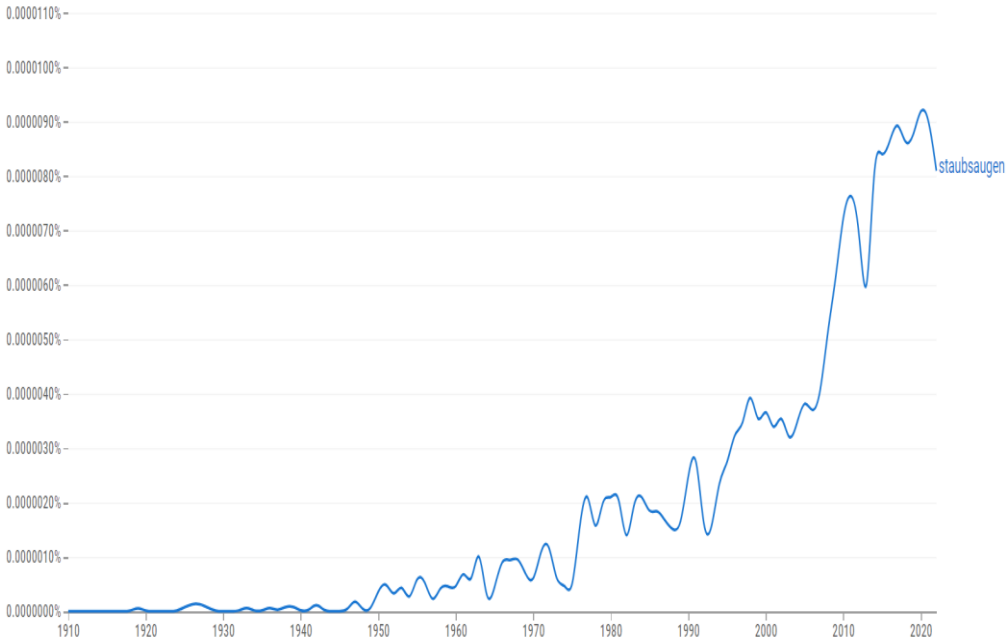

Figure 2: Ngram representation (Google Ngram) of the first usage of the word *staubsaugen* in written sources in the German language

3.3.2 G2: Grammatical items

G2 served as a baseline condition and consisted of utterances that are both structurally well-formed and contextually unmarked. These stimuli were selected to represent standard, uncontroversial usage and were expected to receive uniformly

high acceptability ratings. Their function within the study was to provide an empirical control for interpreting variation in the other, more complex item types. Example (4) illustrates a representative case: Example (4) illustrates a typical case: the utterance is embedded in a plausible everyday context and contains no morphosyntactic anomalies or other incongruities. As such, it allows for the validation of participants' and the model's ability to identify and rate prototypically grammatical sentences as conforming to core structural norms of the language:

(4) Ein Wissenschaftler berichtet über seine Forschungsergebnisse. Er sagt:
'A scientist reports on his research findings. He says:'

Die	Studie	zeigt,		dass	regelmäßiger
DEF.NOM.SG.F	study	show-IND.PRS.3SG		that	regular-NOM.SG.M.S
Schlafmangel		die	kognitive		
sleep-deprivation		DEF.ACC.SG.F	cognitive-ACC.SG.F.W		
Leistungsfähigkeit		beeinträchtigt		und	langfristige
performance		impair-IND.PRS.3SG		and	long-term-ACC.PL
Auswirkungen		auf	die		Gesundheit
effect-ACC.PL		on	DEF.ACC.SG.F		health
haben	kann.				
have-INF	can-IND.PRS.3SG				

'The study shows that regular sleep deprivation impairs cognitive performance and can have long-term effects on health.'

3.3.3 G3: Ungrammatical items

The items in G3 were specifically designed to violate core morphosyntactic constraints of German grammar, irrespective of contextual framing. To isolate grammatical deviance as the source of unacceptability, each item was embedded in a pragmatically coherent context that avoided implausible or confounding discourse effects. The aim of this condition was to assess the extent to which both human participants and the language model detect and penalize structural violations in their acceptability judgments. Example (5) illustrates a representative instance (the ungrammaticality of the sentence is not marked with an asterisk [*] here in order to maintain typographic consistency across examples):

(5) Ein Historiker beschreibt die Französische Revolution. Er sagt:
'A historian explains the French Revolution. He says:'

Die	Revolution	zu	viele	Änderungen
DEF.NOM.SG.F	revolution	to	many-NOM/ACC.PL	change-PL
in	Europa	führte.		
in	Europe	lead-IND.PST.3SG		

Here, the sentence, a declarative main clause, exhibits two major morphosyntactic violations. First, it displays an unlicensed SOV linearization, placing the finite verb *führte* in final position, as would be expected in a subordinate clause introduced by a complementizer rather than in the second position required by the V2 constraint. Second, the quantified phrase *zu viele Änderungen* introduces a case conflict: the preposition *zu* governs the dative, but the form *viele* reflects nominative or accusative morphology, rather than the correct dative *vielen*. The target form would be: *Die Revolution führte zu vielen Änderungen*.

3.3.4 G4: Diatopically marked items

The items in G4 were selected to test evaluators' sensitivity to geographically anchored variation in morphosyntax and at the morpholexical interface. The regional homogeneity of the human sample – all participants were recruited from a Western German region – provided a controlled baseline allowing for a focused investigation into how specific diatopic features are perceived by speakers with broadly comparable linguistic backgrounds. Importantly, the stimuli in this group did not include utterances in full dialect, but rather sentences formulated in colloquial regional German. Each item was designed to feature a morphosyntactic or lexical element that is unambiguously associated with a specific geographic area of the German-speaking world, while still remaining close to informal standard usage. Example (6) illustrates one such item:

(6) Maria spricht über die Zeit nach der Scheidung eurer gemeinsamen Freundin Juliane. Sie sagt:
'Maria is talking about the time after your mutual friend Juliane's divorce. She says:'

Da	hat	sie	sich	nur	schwer
there	AUX.IND.PRS.3SG	she-NOM	REFL.ACC	only	hard
von	erholt,	aber	jetzt	geht	es	besser.
from	recover-PTCP	but	now	go-IND.PRS.3SG	it-NOM	better

'She had a hard time recovering from it, but now she's doing better.'

The feature exemplified in (6) is the optional splitting of prepositional adverbs such as *davon* ('from that'), *dafür* ('for that'), *damit* ('with that') and *dagegen* ('against that'), a phenomenon characteristic of Western and Northern varieties of German. In other regional varieties and dialects across the German-speaking area, such splitting is either ruled out or allowed only under more restrictive conditions (Fleischer 2002a, 2002b; Negele 2012; Leser 2012; Cirkel & Freywald 2021). These adverbs are morphologically complex, consisting of a deictic *da-* element (historically derived from a locative or temporal *d*-adverb) and a preposition. In the example under discussion, *davon* ('from it', here referring to the divorce) appears

in a split configuration: the *da-* element is fronted, while the prepositional part surfaces preverbally in the middle field. This construction is generally judged fully acceptable only by speakers from Western or Northern regions. In Standard German, by contrast, the two components remain adjacent (*Davon hat sie sich nur schwer erholt* or *Sie hat sich nur schwer davon erholt*).

3.3.5 G5: Diastratically / diaphasically marked items

This group targeted constructions that, while not grammatically deviant, display pronounced (in)formal or group-specific features. Crucially, the focus here was not on structural well-formedness per se, but on the appropriateness of register in relation to the discourse context. While the primary analytical perspective adopted here is diastratic in that the relevant forms are associated with particular social and group-specific varieties, the resulting register effects can in specific interactional settings also be understood diaphasically, i.e. as arising from a mismatch between situational context and stylistic choice. To systematically probe this dimension, the group was internally balanced: six items were embedded in contexts where the register was pragmatically congruent, while the remaining six featured the same linguistic forms in settings where they were register-inappropriate or stylistically incongruous. All test items remained within the bounds of colloquial standard German. (7) illustrates a representative case from this group:

(7) Ein Universitätsprofessor hält eine formelle Vorlesung über klassische Philosophie. Er sagt:
‘A university professor is giving a formal lecture on classical philosophy. He says:’

Der	Kant	hat’s		damals	ja
DEF.NOM.SG.M	Kant	AUX.IND.PRS.3SG-IT.ACC		back-then	PRT
voll	drauf	gehabt,	ey!		
totally	on-it	PTCP-have-PTCP	PRT		

‘Kant really nailed it back then, huh!’

This example illustrates a prototypical case of register mismatch: the utterance is strongly marked as informal and youth-coded for different reasons: the use of a definite article to accompany the proper name *Kant*, the use of the expression *es voll drauf haben* ‘nail it’, which is very low-register and associated with colloquial or youth sociolects, and the sentence-final particle *ey*, which functions as a discourse marker typical of very informal orality, particularly in youth and urban varieties. Its use in the highly formal setting of an academic lecture violates expectations associated with professional register and discourse norms. The item is thus predicted to receive low acceptability ratings, not due to grammatical anomaly, but due to its contextual inappropriateness.

In the following section, the results of the study are presented and discussed comparatively.

4 Results and discussion

This section presents the comparative results of the two evaluation conditions described above: S1, which collected acceptability judgments from native speakers of German, and S2, in which the test items were submitted to ChatGPT-4 in a single-pass prompt. Each stimulus was evaluated once by the AI model, resulting in a single scalar rating per item. In contrast, human participants rated each sentence independently and for the purposes of this analysis, the mean value of the 65 ratings per item was computed. All statistical comparisons below are based on these averaged human scores vis-à-vis the corresponding single output generated by ChatGPT-4. This asymmetry reflects the fundamentally different nature of the two evaluators and necessitates the use of appropriate statistical procedures (e.g., Welch's t-test) to assess the significance of observed differences. The analyses were carried out using JASP (JASP Team 2024, https://jasp-stats.org/), an open-source software package for statistical inference. Overall, the results do not provide evidence in favor of H_1; instead, they support H_0, indicating that no statistically significant differences were found between the grammaticality and acceptability judgments generated by ChatGPT-4 and those of native German speakers. Within the human participant group, additional analyses revealed no statistically significant effects of gender or level of academic qualification (BA vs. MA) on acceptability judgments. As these variables did not contribute systematically to the observed patterns, they are not considered further in the discussion that follows.

4.1 Group-level variation: ANOVA results

To evaluate whether the five stimulus groups (G1 through G5) elicited systematically different acceptability ratings, a one-way analysis of variance (ANOVA) was conducted.[8] This test was applied separately to the two datasets – S1 (human participants) and S2 (ChatGPT-4) – as well as to the combined dataset, in order to assess both within-system differentiation and cross-system coherence.

In S1, the ANOVA yielded a highly significant main effect of stimulus group on acceptability ratings ($p < .001$), with a large effect size ($\eta^2 = 0.484$). This indicates that human participants reliably distinguished among the five groups,

[8] The one-way ANOVA is a statistical procedure used to test whether the mean values of a continuous variable differ significantly across multiple independent groups. It is particularly appropriate in the present context because it allows for the assessment of whether the five stimulus categories G1-G5, which were deliberately constructed to reflect distinct linguistic properties, elicited systematically different acceptability judgments. Rather than conducting multiple pairwise comparisons, ANOVA provides a single overall test for group-level effects, helping to identify whether any reliable differences emerge across the set of ratings.

assigning differentiated scores in accordance with the grammatical, pragmatic and sociolinguistic properties of the stimuli. This result confirms that participants responded sensitively to the intended experimental manipulations and validates the internal structure of the stimulus set.

In S2, the model-generated responses revealed a strikingly similar pattern. The ANOVA performed on ChatGPT-4's ratings also returned a highly significant main effect ($p < .001$), with a comparable effect size ($\eta^2 = 0.474$). This demonstrates that the model, like human speakers, is sensitive to the systematic contrasts built into the test material and applies graded judgments across the five stimulus categories.

When the two datasets were pooled and analyzed jointly, the ANOVA again yielded a highly significant main effect ($p < .001$), with a large effect size ($\eta^2 = 0.463$). This reinforces the conclusion that the group structure of the stimulus set was robustly detected by both evaluators and affirms the external validity of the experimental design:

Table 2: One-way ANOVA results for S1, S2 and combined data.

Condition	p-value	η^2 (effect size)	Interpretation
S1	< 0.001	0.484	Significant group differences; large effect size
S2	< 0.001	0.474	Significant group differences; large effect size
combined	< 0.001	0.463	Significant group differences across both systems; large effect size

Importantly, these group-level differences were not only statistically significant but also *theoretically expected*. The stimulus groups were deliberately constructed to probe distinct linguistic dimensions – ranging from fully grammatical sentences to structurally deviant, regionally marked or register-sensitive utterances. The emergence of strong inter-group contrasts confirms that both human speakers and the model responded in accordance with the intended typological distinctions. In other words, the observed differences exclude random variability and provide empirical support for the study's design rationale.

Despite this convergence, two points of divergence merit attention. First, the rating distribution among human participants exhibited greater variability, as indicated by higher standard deviations and standard errors. This is consistent with previous research showing that speaker intuitions tend to be gradient, context-sensitive and heterogeneous. Second, ChatGPT-4 displayed zero variance in its ratings of grammatical control items (G2), assigning a uniform score of 5 across all such items. This categorical treatment of grammaticality points to a reduced sensitivity to subtle contextual or prosodic cues that may shape human judgments, even when sentences are structurally well-formed.

4.2 Study-wise and group-wise comparisons: Welch's T-Tests

To investigate the extent and significance of cross-system differences in acceptability judgments, a series of independent Welch two-sample t-tests was

conducted.[9] These comparisons were carried out (i) individually for each of the five stimulus groups (G1-G5), in order to identify localized divergences between the two evaluators, and (ii) globally across the full dataset, pooling all 60 items. While the group-specific analyses offer a more fine-grained view of category-level variation, the global test – reported at the end of this section – serves to complement these findings by assessing the overall alignment between ChatGPT-4 and the human participants across the full range of stimuli.

4.2.1 Results for G1 (grey-zone items)

For G1, the Welch two-sample t-test revealed no statistically significant difference between the acceptability ratings assigned by the human participants (S1) and ChatGPT-4 (S2), with $t(20.779) = 1.065$, $p = 0.299$. As shown in Figure 3, the mean score in the model condition was 2.917 (SD = 1.621), whereas the human sample yielded a slightly higher mean of 3.549 (SD = 1.266). Despite this numerical discrepancy, the difference remains statistically non-significant, indicating that both evaluators responded comparably to the stimuli in this group. In Figures 3-7, generated using JASP, the German terms *Bewertung* (lit. 'judgment') and *Quelle* (lit. 'source') denote, respectively, the dependent variable (the ratings) and the independent variable or grouping factor, i.e., the variable that defines the distinction between S1 and S2:

Independent Samples T-Test ▼

Independent Samples T-Test ▼

	t	df	p
Bewertung	1.065	20.779	0.299

Note. Welch's t-test.

Descriptives

Group Descriptives

	Group	N	Mean	SD	SE	Coefficient of variation
Bewertung	S1	12	3.549	1.266	0.365	0.357
	S2	12	2.917	1.621	0.468	0.556

Figure 3: G1: mean ratings, standard deviations and Welch's t-test results (S1 vs. S2).

[9] The Welch two-sample t-test is a statistical method used to compare the mean values of two independent groups, even when those groups have unequal variances or sample sizes. In the context of this study, it was employed to assess whether the acceptability ratings provided by ChatGPT-4 and human participants differed significantly across individual stimulus groups and at the overall dataset level. The Welch variant was chosen over the standard t-test because it offers a more robust estimate when the assumption of equal variance across groups is not guaranteed – an important consideration, given the different sources and rating behaviors involved.

The elevated standard deviation in ChatGPT-4's responses, as well as the higher coefficient of variation (0.556 for S2 vs. 0.357 for S1), suggests a somewhat more dispersed response pattern on the part of the model. Nevertheless, in the absence of statistical significance, these differences do not support the inference of a systematic divergence in how grey-zone items were evaluated across the two study conditions.

4.2.2 Results for G2 (grammatical items)

In G2, no statistical comparison could be performed due to the absence of variance in the model's ratings. As indicated in Figure 4, ChatGPT-4 (S2) assigned the maximum score of 5.000 (5 = very natural) to all 12 items, resulting in a standard deviation of 0.000. By contrast, the human participants (S1) showed a slightly lower mean rating of 4.673 (SD = 0.201), indicating a small degree of individual variability in their evaluations.

Because the model's ratings lacked any variance, a Welch t-test could not be calculated, as statistical significance testing requires variation within both groups.[10] Nevertheless, the descriptive data clearly show that ChatGPT-4 adopted a uniformly categorical stance towards this stimulus group, treating all grammatical control items as fully acceptable without exception. Human ratings, while also very high on average, allowed for minimal gradience, possibly reflecting slight contextual or prosodic effects not factored into the model's evaluation.
This result confirms that both evaluators recognized the canonical well-formedness of the G2 items, though ChatGPT-4 appears more rigid in its application of grammaticality criteria:

Independent Samples T-Test ▼

Independent Samples T-Test ▼

	t	df	p
Bewertung	NaN[a]		

Note. Welch's t-test.
[a] The variance in Bewertung is equal to 0 after grouping on Quelle

Descriptives

Group Descriptives

	Group	N	Mean	SD	SE	Coefficient of variation
Bewertung	S1	12	4.673	0.201	0.058	0.043
	S2	12	5.000	0.000	0.000	0.000

Figure 4: G2: mean ratings, standard deviations and Welch's t-test results (S1 vs. S2).

[10] Because the *t*-test assesses differences in means relative to within-group variability, it cannot operate when one group's variance is zero.

Although a formal significance test could not be computed for this group, the descriptive pattern is nonetheless revealing. ChatGPT-4 (S2) consistently assigned the maximum score of 5.0 to every item in this group, resulting in a complete absence of variance. This categorical response behavior suggests that the model evaluates grammaticality in binary terms: if a sentence is structurally well-formed and contextually plausible, it is rated as fully acceptable, without gradience or contextual differentiation.

In contrast, the human participants (S1) exhibited slightly lower ratings on average (M = 4.673), accompanied by a small but non-zero standard deviation. This indicates a more conservative evaluation strategy, with speakers allowing for subtle variation even within the domain of grammatical sentences, which likely reflects pragmatic expectations or individual interpretive thresholds.

Overall, his pattern reveals an important divergence in rating behavior: ChatGPT-4 appears to treat grammaticality in more absolute terms, assigning consistently perfect scores to structurally canonical items. Human participants, on the other hand, show a slightly more measured response, even when evaluating clearly well-formed sentences. Their lower average scores and minimal variance suggest a more cautious and context-sensitive approach, possibly shaped by fine-grained intuitions about context or discourse fit. While both systems ultimately agree on the grammaticality of the stimuli, human judgments reflect a nuanced sensitivity that the model does not replicate.

4.2.3 Results for G3 (ungrammatical items)

For G3, the comparison between S1 and S2 did not yield a statistically significant difference. The Welch two-sample *t*-test produced a *t*-value of 0.557 with 19.464 degrees of freedom and a *p*-value of 0.584. This clearly exceeds the conventional threshold of $\alpha = 0.05$, indicating that the difference between the two groups' mean ratings is not statistically meaningful.

As shown in Figure 5, the human participants assigned an average score of 1.242 (SD = 0.267), while ChatGPT-4's ratings averaged slightly lower at 1.167 (SD = 0.389). Both values lie in the lower segment of the acceptability scale, confirming that both evaluators consistently recognized the ungrammatical nature of the items in this group. The model's marginally lower average suggests a slightly more lenient rating pattern, though this difference does not reach significance.

Independent Samples T-Test

Independent Samples T-Test ▼

	t	df	p
Bewertung	0.557	19.464	0.584

Note. Welch's t-test.

Descriptives

Group Descriptives

	Group	N	Mean	SD	SE	Coefficient of variation
Bewertung	S1	12	1.242	0.267	0.077	0.215
	S2	12	1.167	0.389	0.112	0.334

Figure 5: G3: mean ratings, standard deviations and Welch's t-test results (S1 vs. S2).

The difference in standard deviations and coefficients of variation (0.215 for humans vs. 0.334 for the model) further indicates that ChatGPT-4 exhibited more variability across individual items, possibly reflecting fluctuating confidence in syntactic acceptability. However, given the non-significant p-value, this should not be interpreted as evidence for a systematic divergence between the two systems in the evaluation of G3 items.

4.2.4 Results for G4 (diatopically marked items)

Also in the case of G4, the Welch two-sample t-test revealed no statistically significant difference between the acceptability ratings of S1 and those of S2. The test yielded a t-value of -0.970 with 21.998 degrees of freedom and a corresponding p-value of 0.342, clearly exceeding the significance threshold of $\alpha = 0.05$.
Figure 6 shows that ChatGPT-4 produced a slightly higher average acceptability rating (M = 3.500, SD = 1.508) compared to the human participants (M = 2.900, SD = 1.522). Despite this numerical discrepancy of 0.6 scale points, the difference does not reach statistical significance and should thus not be interpreted as reflecting a systematic divergence in judgment:

Independent Samples T-Test ▼

Independent Samples T-Test

	t	df	p
Bewertung	−0.970	21.998	0.342

Note. Welch's t-test.

Descriptives ▼

Group Descriptives

	Group	N	Mean	SD	SE	Coefficient of variation
Bewertung	S1	12	2.900	1.522	0.439	0.525
	S2	12	3.500	1.508	0.435	0.431

Figure 6: G4: mean ratings, standard deviations and Welch's t-test results (S1 vs. S2).

The relative similarity in the standard deviations and coefficients of variation between the two groups (0.525 for S1, 0.431 for S2) confirms that both the human participants and ChatGPT-4 displayed comparable levels of intra-group variability in their evaluations of regionally marked constructions. While the slightly more favorable ratings from ChatGPT-4 may hint at a lower sensitivity to diatopic variation (possibly due to its exposure to a wide spectrum of dialectal data during training) these differences remain statistically non-significant and thus cannot be generalized without further evidence.

4.2.5 Results for G5 (diastratically / diaphasically marked items)

For G5, the Welch two-sample t-test revealed no statistically significant difference between the ratings provided by the human participants and ChatGPT-4. The test produced a t-value of 0.021 with 21.626 degrees of freedom, and a p-value of 0.983 – well above the conventional alpha threshold of 0.05. This outcome clearly indicates the absence of any meaningful statistical divergence between the two conditions.

As shown in Figure 7, the mean rating for S1 was 3.100 (SD = 1.770), while S2 yielded a very similar mean of 3.083 (SD = 2.021). The almost identical average scores suggest a shared evaluation of the register- or socially marked expressions tested in this group. In both cases, the responses display high variability, reflected in relatively large standard deviations and coefficients of variation (0.571 for S1, 0.655 for S2), indicating considerable within-group fluctuation in how these socially colored constructions were judged:

Independent Samples T-Test ▼

Independent Samples T-Test ▼

	t	df	p
Bewertung	0.021	21.626	0.983

Note. Welch's t-test.

Descriptives

Group Descriptives

	Group	N	Mean	SD	SE	Coefficient of variation
Bewertung	S1	12	3.100	1.770	0.511	0.571
	S2	12	3.083	2.021	0.583	0.655

Figure 7: G5: mean ratings, standard deviations and Welch's t-test results (S1 vs. S2).

Taken together, these results suggest that ChatGPT-4 and human participants exhibit closely aligned intuitions when it comes to stylistically or socially marked constructions. The convergence in mean values, coupled with similar patterns of dispersion, implies that both evaluation systems are comparably sensitive to the pragmatic appropriateness of diastratic / diaphasic variation, at least under the conditions established by the experimental design.

To provide a concise overview of the group-wise comparisons discussed above, Table 3 summarizes the statistical significance of the Welch two-sample t-tests across the five stimulus categories. This synopsis is intended to help the reader contextualize the outcome of the individual analyses before turning to the global evaluation presented in Section 4.2.6:

Table 3: Overview of group-wise statistical comparisons between S1 and S2.

Group	**Stimulus type**	**Welch t-test result**	**Statistical significance**
G1	grey-zone items	$p = 0.299$	not significant
G2	grammatical items	—	not testable (S2: SD = 0)
G3	ungrammatical items	$p = 0.584$	not significant
G4	diatopically marked items	$p = 0.342$	not significant
G5	diastratically / diaphasically marked items	$p = 0.983$	not significant

4.2.6 Global comparison across all items

As can be inferred from the results reported for the individual stimulus groups, no systematic (that is, statistically significant) divergence emerges between the overall response patterns of ChatGPT-4 and the human participants. Nonetheless, to verify this impression and ensure statistical completeness, a global Welch two-sample t-test was conducted on the full set of 60 items, pooling across all stimulus groups (G1-G5). The test yielded a t-value of -0.094 with approximately 115.74 degrees of freedom and a corresponding p-value of 0.925. This result clearly fails to reach

statistical significance, indicating that the overall acceptability ratings produced by the two evaluators do not differ in any meaningful way.

Although the group-specific comparisons already revealed certain localized divergences in mean scores, this aggregated analysis confirms that – when all items are considered together – ChatGPT-4 and the human participants display a high degree of global alignment in their acceptability assessments. The absence of a significant difference highlights the extent to which the model mirrors native-speaker judgments in broad terms. It thereby establishes a robust baseline for interpreting the more fine-grained contrasts observed at the group level and affirms the model's general sensitivity to linguistic well-formedness within the confines of the present experimental design.

5 Selected edge cases and qualitative insights

While the statistical analyses presented in Section 4 demonstrate a general convergence between ChatGPT-4 and the human control group, not all stimulus groups warrant equal attention from a qualitative perspective. The evaluations of G2 (grammatical items), G3 (ungrammatical items) and G5 (diastratically / diaphasically marked items) yielded no particularly notable divergences or irregularities: both evaluators produced relatively consistent judgments and the distribution of ratings exhibited no meaningful anomalies. G1 (grey-zone items), by contrast, contains a number of individual outliers that – despite falling short of statistical significance – suggest minor deviations in the way ambiguous or structurally marginal constructions are assessed by the model as compared to human speakers. Most salient, however, is G4 (diatopically marked items), where the absence of statistically significant differences appears to mask a systematic evaluative tendency in ChatGPT-4's responses. This section turns to a more detailed qualitative analysis, focusing on selected examples that shed light on the sources of variation and the interpretive behavior of the model.

5.1 Edge cases in G1

5.1.1 Isolated deviations in the model's treatment of grey-zone items

Within the grey-zone stimulus set (G1), three items – stimuli 8, 9 and 11 – stand out for exhibiting more or less pronounced rating asymmetries between ChatGPT-4 and the human control group. Although these discrepancies are neither numerous nor consistent enough to result in statistical significance at the group level, they nonetheless reveal a qualitatively meaningful divergence in evaluation patterns. In all three cases, ChatGPT-4 assigns substantially lower acceptability ratings (ranging from 2 to 4), whereas the human participants respond with uniformly high scores ($M > 4.1$). The relevant stimulus examples are presented below. All three items illustrate the phenomenon of non-V2 verbs discussed in Section 3.1.1. Item 8 is particularly notable: despite featuring a verb that is both attested and lexically

established, it receives only moderate acceptability (3/5) from the model. The verb *kopfstehen* ('to be upside down') is listed in authoritative lexical resources such as the DWDS (*Digitales Wörterbuch der deutschen Sprache*; dwds.de/), where it appears in both separated and V2 constructions—for example: *Das ganze Haus steht vor Aufregung kopf* ('The whole house is upside down with excitement'; https://www.dwds.de/wb/kopfstehen). The infinitival form used in Item 8, *kopfzustehen*, clearly signals this type of separable construction, with the marker *zu* intervening between the prefix *kopf* ('head') and the verb stem *steh-* ('stand'):[11]

(8) Item 8
Sie befinden sich in einem Teammeeting, in dem eine unerwartete und drastische Änderung der Projektanforderungen bekannt gegeben wurde. Die Neuigkeiten haben alle im Team überrascht und die bisherigen Pläne über den Haufen geworfen. Ihre Kollegen sind verunsichert und diskutieren, wie sie mit der neuen Situation umgehen sollen. Einer von ihnen sagt:
'You are in a team meeting where an unexpected and drastic change to the project requirements has just been announced. The news has taken everyone on the team by surprise and thrown the previous plans into disarray. Your colleagues are unsettled and are discussing how to deal with the new situation. One of them says:'

Seit	wir	die	neuen	Anforderungen		
since	we-NOM	DEF.ACC.PL	new-ACC.PL.W	requirement-ACC.PL		
erhalten	haben,		scheint	die		
receive-PTCP	AUX.IND.PRS.1PL		seem-IND.PRS.3SG	DEF.NOM.SG.F		
ganze	Welt	plötzlich	kopfzustehen!			
whole-NOM.SG.F.W	world	suddenly	head-to-stand.INF			

'Ever since we received the new requirements, it feels like the whole world has suddenly been turned upside down!'

[11] Note that in German, the particle *zu* functions as an obligatory marker of infinitival status in certain non-finite constructions. In sentences like (i), where the lexical verb is simplex (e.g., *red-* 'to talk'), *zu* directly precedes the verb:

(i)	Maria	hat	keine	Lust,	mit	Hans	zu reden.
	Maria	have-IND.PRS.3SG	no-ACC.SG.F	desire	with	Hans	to talk-INF
	'Maria doesn't feel like talking to Hans.'						

In verbal compounds or pseudo-compounds involving a separable prefix (i.e., a particle that detaches in V2 contexts), *zu* appears between the separable particle and the verbal root, as in (ii):

(ii)	Maria	hat	keine	Lust,	mit	Hans	auszugehen.
	Maria	have-IND.PRS.3SG	no-ACC.SG.F	desire	with	Hans	V.PRT-to-go-INF
	'Maria doesn't feel like going out with Hans.'						

By contrast, when the verb is derived with an inseparable prefix, *zu* cannot intervene morphologically. Instead, it is positioned externally, preceding the entire verb complex, as in (iii):

(iii)	Maria	hat	keine	Lust,	mit	Hans	zu verhandeln.
	Maria	have-IND.PRS.3SG	no-ACC.SG.F	desire	with	Hans	to negotiate-INF
	'Maria doesn't feel like negotiating with Hans.'						

Items 9 and 11 both received a rating of 2/5 (“rather unnatural”) from the AI model, in contrast to average ratings above 4 in S1. Item 9, shown in (9), features the verb *notlanden* (‘to make an emergency landing’) in a configuration where the internal structure of this separable verb compound is interrupted by the infinitival particle *zu*, as in Item 8. Lexical resources such as the DWDS (https://www.dwds.de/wb/notlanden) and Duden (https://www.duden.de/rechtschreibung/notlanden) generally confirm that non-finite, clause-final occurrences of *notlanden* are attested and that the verb is separable in this context. This is in line with findings from previous empirical studies (e.g., Åsdahl-Holmberg 1976: 31-32; Forche 2020: 209) and compatible with the structure used in the corresponding questionnaire item, where the verb appears at the end of the clause with *zu* intervening between the first element and the verb stem:

(9) Item 9
In den Nachrichten wird erzählt, dass während eines Flugs ein medizinischer Notfall auftritt, der eine Notlandung erforderlich macht. Die Journalistin sagt:
‘The news report states that a medical emergency occurred during a flight, requiring an emergency landing. The journalist says:’

Der	Pilot	entschied	sich,	das	
DEF.NOM.SG.M	pilot	decide-IND.PST.3SG	REFL.ACC	DEF.ACC.SG.N	
Flugzeug	wegen	eines	medizinischen		
airplane	due-to	INDEF.GEN.SG.M	medical-GEN.SG.M.W		
Notfalls	an	Bord	notzulanden.		
emergency-GEN.SG	on	board	emergency-to-land-INF		

‘The pilot decided to emergency-land the plane due to a medical emergency on board.’

Item 11 features the verb *brustschwimmen* (‘to swim breaststroke’), which receives a markedly more restrictive lexical treatment in the reference sources. In contemporary usage, the verb is typically attested only in the infinitival form. Where other forms occur – such as the indicative *Perfekt* construction employed in the stimulus (which, at least morphologically, parallels the English present perfect) –, the past participle is formed by inserting the prefix *ge-* between the initial element (*brust* ‘chest’) and the allomorphic verb stem (*-schwomm-*) and is then syntactically combined with the auxiliary *sein* (‘be’). This configuration closely parallels the *zu*-intervention pattern observed in earlier examples (cf. fn. 11).[12] See, for instance,

[12] This can be shown with examples similar to those in fn. 11. Note, however, that unlike *zu*, which is a free morpheme, *ge-* functions as a bound morpheme that is attached to the left of the stem/root in the formation of the past participle of simplex verbs. By contrast, it is not spelled out at all in the case of verbs with a derivational prefix. In (iii), the symbol ‘∅’ indicates that no overt morpheme appears in that position:

the entries in the DWDS (https://www.dwds.de/wb/brustschwimmen) and Duden (https://www.duden.de/rechtschreibung/brustschwimmen):

(10) Item 11
Sie sprechen mit jemandem über Ihre körperliche Fitness und Ihre Fähigkeiten im Schwimmsport. Er fragt nach dem Grund für Ihre gute Leistung oder spezielle Fähigkeit im Schwimmen und Sie erklären, dass Ihre Technik oder Trainingseinheit einen Unterschied gemacht hat. Sie sagen:
'You're talking to someone about your physical fitness or your abilities in swimming. They ask about the reason for your good performance or a particular skill in swimming and you explain that your technique or training made a difference. You say:

Das		liegt	wohl	daran,	dass	ich	als
that-NOM.SG.N		lie-IND.PRS.3SG	PRT	at-it	that	I-NOM	as
Kind	so	oft	brustgeschwommen		bin.		
child	so	often	breast-PTCP-swim-PTCP		AUX.IND.PRS.1SG		

'It's probably because I did so much breaststroke as a child.'

Table 4 summarizes the corresponding ratings: the first column (S1) displays the mean acceptability scores from the human participants, while the second column (S2) lists the single-point ratings assigned by ChatGPT-4:

Table 4. Selected G1 items with marked rating divergence between S1 and S2.

	S1 (human participants)	S2 (ChatGPT-4)
Item 8	4.35	3
Item 9	4.65	2
Item 11	4.17	2

An explanation for the rating divergences observed in Items 8, 9 and 11 must take into account both the specific morphosyntactic structure of the stimuli and the fundamentally different architectures – cognitive in the case of humans, computational in the case of ChatGPT-4 – that govern linguistic evaluation. All three items involve German separable verb compounds occurring in non-finite or participial forms, where the verbal complex is disrupted by the insertion of the infinitival marker *zu* or the inflectional prefix *ge-*. These constructions are attested,

(i) Maria hat nicht mit Hans geredet.
Maria AUX.IND.PRS.3SG NEG with Hans PTCP-talk-PTCP
'Maria did not talk to Hans.'

(ii) Maria ist nicht mit Hans ausgegangen.
Maria AUX.IND.PRS.3SG NEG with Hans V.PRT-PTCP-go-PTCP
'Maria did not go out with Hans.'

(iii) Maria hat nicht mit Hans Ø-verhandelt.
Maria AUX.IND.PRS.3SG NEG with Hans negotiate-PTCP
'Maria did negotiate with Hans.'

internally compositional and derivable from productive morphological rules. However, their surface forms are highly marked, lexically specific and rare in standard usage. They occupy a grey zone in the linguistic system not because they violate any structural constraints, but because they exist at the intersection of grammatical possibility and uncertain distributional acceptability. Native speakers may hesitate, vary in their preferences, or even self-correct when confronted with such forms in spontaneous speech. Nevertheless, as the results show, in controlled settings like acceptability judgment tasks, speakers tend to recognize these forms as structurally licit, drawing on abstract knowledge of separable verbs and productive inflectional patterns. ChatGPT-4, by contrast, rates all three stimuli significantly lower than the human baseline, revealing not just a performance divergence, but a difference in evaluative foundations.

To understand the underlying contrast more precisely, we must consider the nature of the linguistic data available to each system. Human participants bring to the task a deeply entrenched linguistic competence shaped by prolonged, situated exposure to naturalistic input. Their internal grammar is the product of early and sustained acquisition, supported by biologically evolved, language-specific cognitive mechanisms. This input is not only syntactically rich and pragmatically embedded, but also filtered over time through a balance of positive evidence, indirect negative evidence, and context-sensitive reinforcement. Even when confronted with low-frequency or structurally marked forms, speakers are able to rely on analogical reasoning, morphological generalization, and intuitive judgments rooted in what we might call a "probabilistic-but-structured" mental grammar.

Crucially, this capacity should not be understood merely as a manifestation of a general cognitive ability to accommodate unusual or unfamiliar stimuli. Rather, it reflects a form of tacit, language-specific knowledge that allows speakers to distinguish between rare but rule-conforming structures and configurations that violate the grammatical constraints of their language. Therefore, while human judgments are not always confident or uniform (especially in marginal cases such as those under study here), speakers are nonetheless able to recognize unfamiliar constructions as potential outputs of their grammatical system precisely because these constructions respect the morphosyntactic principles internalized in the course of acquisition. By contrast, forms that violate these principles, even if superficially similar or equally novel, are reliably rejected. In this sense, speakers' judgments do not merely reflect cognitive flexibility in the face of novelty, but a specifically linguistic competence that is sensitive to the abstract rules governing well-formedness in a given language.

Accordingly, their judgments reflect a cognitive architecture that combines adaptability with constraint, supporting both fluidity and resilience in the face of lexical sparsity and syntactic irregularity, while remaining anchored in language-specific grammatical knowledge. ChatGPT-4, by contrast, is trained not on curated linguistic input designed to support language acquisition, but on massive amounts of web-scraped text, a corpus that is both extensive and noisy. These corpora are

skewed toward high-frequency, standard, and contextually conventionalized usage patterns and tend to underrepresent syntactically complex, morphologically marked, or stylistically restricted constructions. Moreover, forms that do occur in formal lexical resources such as DWDS or Duden may not appear frequently enough in running text to register as "grammatical" in the model's statistical expectations. Since the model's training objective is next-token prediction rather than rule induction, it has no way of distinguishing productive-but-rare from ungrammatical-but-frequent: both are collapsed into relative likelihoods over string continuations. This results in a kind of distributional overfiltering whereby forms that are structurally well-formed but rare are downgraded simply because they do not match high-probability patterns in the training data. The underlying problem, therefore, is not that the model makes errors in specific cases, but that it lacks access to a generative system capable of abstracting away from surface frequency toward underlying structure – or, in the terms of Kahneman (2011) (cf. 3.2), to anything resembling the intuitive, fast-acting evaluations characteristic of System 1.

5.1.2 Averaging without alignment

A particularly instructive case of evaluative asymmetry arises in Item 3 of the stimulus set (reproduced in (11) from (3b) above), which features the verb *staubsaugen* ('to vacuum') in its finite, V2-separated form:

(11)	Ich	sauge	gerade	den	Teppich	staub.
	I-NOM	suck-IND.PRS.1SG	right-now	DEF.ACC.SG.M	carpet	dust
	'I'm just vacuuming the carpet.'					

The sentence follows the canonical word order of separable verbs in German main clauses: the inflected verbal head *sauge* ('suck') occupies the V2 slot, while the separable modifier *staub* ('dust') surfaces clause-finally. Ratings from human participants reveal a distinctly polarized distribution: responses cluster at opposite ends, with high frequencies for both 'very unnatural' (1) and 'rather natural' (4), producing a bimodal profile indicative of interspeaker variability and possibly competing stylistic or register-based intuitions:

Table 5. Ratings in S1 for Item 3 (*Ich sauge gerade den Teppich staub*).

Answer	Responses	Percentage	Mean acceptability rating
1 (‘very unnatural’)	17	26.2%	3.11
2 (‘rather unnatural’)	6	9.2%	
3 (‘moderately natural’)	5	7.7%	
4 (‘rather natural’)	27	41.5%	
5 (‘very natural’)	10	15.4%	

By contrast, when presented with the corresponding inseparable variant (*Ich staubsauge gerade den Teppich*; Item 4), the same participants show a clear preference for naturalness, with the majority opting for ‘rather natural’ or ‘very natural’ (Figure 8). This aligns with the model’s own rating of 5/5 for Item 4:

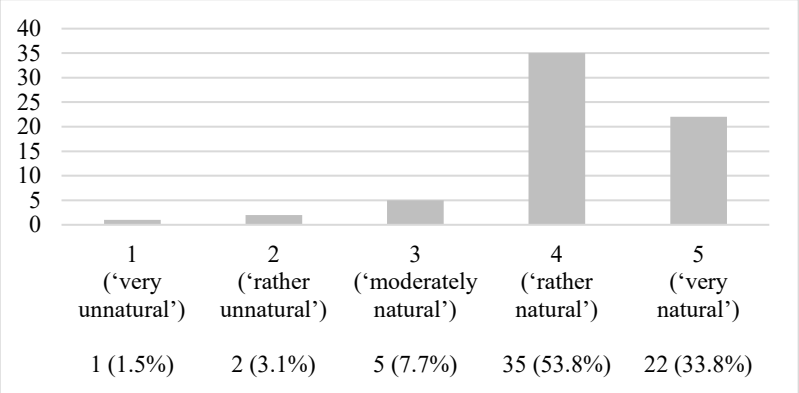

Figure 8: Ratings in S1 for Item 4 (*Ich staubsauge gerade den Teppich*).

As for Item 3, the average acceptability score in S1 is 3.11. Yet the specific rating that corresponds to this mean – ‘moderately natural’ (3/5) – is chosen by only a small minority (7.7%), in fact making it the least represented category in the human distribution. Strikingly, ChatGPT-4 selects precisely this middle value, thereby aligning numerically with the group mean while diverging from the sharply bimodal distribution of actual human judgments. This apparent convergence at the level of the mean thus conceals a fundamental difference in evaluative strategy: whereas human respondents rely on lexical, stylistic, or register-related heuristics that produce polarized judgments, the model appears to interpolate toward a central value, reflecting distributional averaging rather than categorical preference. In this sense, the model’s rating illustrates how numerical alignment with group means can mask a deeper misalignment in evaluative logic.

5.2 A diatopic bias in G4?

A first point of entry into the discussion of possible diatopic effects in ChatGPT-4’s evaluation behavior is provided by Item 41 of the G4 stimulus set, reproduced

in (12). The sentence features a split prepositional adverb typical of western and northern varieties of German (*da ... von* 'of that', see 3.3.4 above):

(12) Tom wird in einem Meeting nach den Details eines Projekts gefragt und gesteht, dass ihm diese Informationen fehlen. Er sagt:
'Tom is asked about the details of a project in a meeting and admits that he doesn't have this information. He says:'

Ich muss ganz ehrlich sagen, dass ich da nichts von weiß.
I-NOM must-IND.PRS.1SG completely honestly say-INF that I-NOM there nothing of know-IND.PRS.1SG
'I have to be honest, I don't know anything about that.'

While the variant *davon* is standard in formal registers and in normative grammar (*dass ich davon nichts weiß* or *dass ich nichts davon weiß*), the split form is nonetheless attested and widely used in regionally spoken varieties. From a structural point of view, both realizations are fully interpretable and syntactically well formed. What distinguishes them is their diatopic markedness – i.e., their association with specific geographical dialect zones within the German-speaking area.

When presented with this stimulus, human participants in S1 responded overwhelmingly positively: as shown in Table 6, 72.3% of the group rated the item as 'very natural' (5/5), with another 20% selecting 'rather natural' (4/5). Only 1.5% rated it as 'rather unnatural' and none of the participants considered the sentence 'very unnatural'. This yields a mean acceptability rating of 4.63, placing the item at the upper end of the naturalness scale. The corresponding distribution is visualized in Figure 9, which clearly shows a rightward skew with the vast majority of judgments clustered at the top of the scale.

Table 6. Ratings in S1 for Item 41 (*Ich muss ganz ehrlich sagen, dass ich da nichts von weiß*).

Answer	**Responses**	**Percentage**	**Mean acceptability rating**
1 ('very unnatural')	0	0.0%	4.63
2 ('rather unnatural')	1	1.5%	
3 ('moderately natural')	4	6.2%	
4 ('rather natural')	13	20.0%	
5 ('very natural')	47	72.3%	

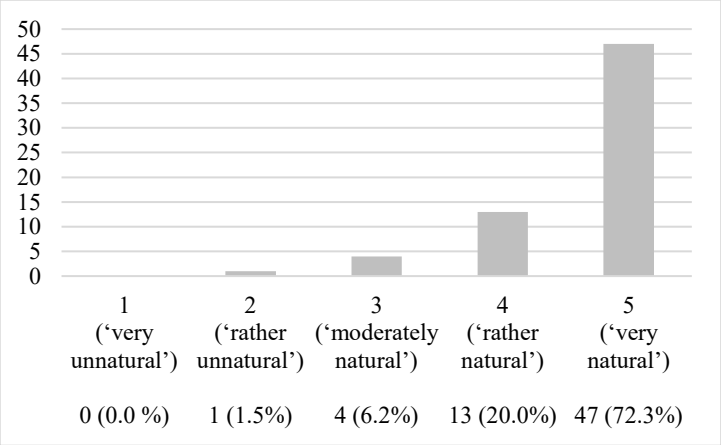

Figure 9. Distribution of S1 ratings for diatopically marked Item 41.

What renders this case particularly salient in the present context is that ChatGPT-4's evaluation closely mirrors the mean human judgment, assigning a rating of 5/5. This numerical alignment gains further interpretive depth when viewed in relation to the linguistic background of the participant cohort. As noted earlier, all human participants in the S1 group are speakers of Western German varieties – precisely those regional dialects in which the split *da*-adverb construction enjoys broad acceptability and frequent usage. Against this backdrop, the model's rating does not appear to reflect an abstract, variety-independent grammatical baseline; rather, it converges with a regionally specific pattern of acceptability. In this respect, the model's behavior in this instance is consistent with that of a Western German speaker.

This observation carries theoretical weight for at least two reasons. First, it can be taken to demonstrate that the model's behavior in this instance is not solely reducible to expectations shaped by dominant or "standard" usage patterns, but may instead plausibly reflect distributional exposure to regionally skewed forms within the broader corpus landscape on which it was trained. Since large-scale language models like ChatGPT-4 are trained on heterogeneous textual corpora, which in all likelihood also include informal dialogues, dialectal features, and transcribed speech, it is plausible that such systems can, under certain conditions, internalize and replicate diatopic variation, at least when the relevant forms are sufficiently represented.

Secondly, the case can be said to establish an important contrastive baseline: if the model's judgment in this instance aligns closely with that of Western German speakers, then systematic misalignment in other, structurally analogous cases (as explored in the following subsection) may reasonably be taken to signal not mere variance, but a deeper distributional bias – one stemming from underrepresentation, overgeneralization or frequency asymmetries in the training data. At first glance, one might take this example as evidence that ChatGPT-4 accepts as grammatical any structure that is formally well-formed, irrespective of the regional variety in

which it occurs. But as will become apparent, this interpretation is arguably too optimistic: the model's performance in such contexts is far from uniform.

Evidence for this emerges from a structurally analogous, but evaluatively contrasting case – Item 39 of the G4 set –, reproduced in (13):

(13) Klaus berichtet seinem Kollegen am Ende eines langen Messetages, dass ihm die Beine vom ständigen Stehen wehtun. Er sagt:
'Klaus tells his colleague at the end of a long day at the trade fair that his legs hurt from standing all the time. He says:'

Heute	bin	ich	die	ganze	
today	AUX.IND.PRS.1SG	I-NOM	DEF.ACC.SG.F	whole-ACC.SG.F.W	
Zeit	gestanden,	mir	tun	die	Beine
time	PTCP-stand-PTCP	I-DAT	do-IND.PRS.3PL	DEF.NOM.PL	leg-NOM.PL
weh!					
sore					

'I've been on my feet all day, my legs are killing me!'

Here, the sentence features a different regionally marked grammatical pattern: the use of the auxiliary *sein* ('be') with the verb *stehen* ('stand') to form the indicative *Perfekt* (*bin gestanden*, lit. 'am stood' = 'have been standing') (cf. 5.1.1). While this auxiliary selection is entirely standard in southern varieties of German, where posture verbs such as *stehen* 'stand', *sitzen* 'sit' and *liegen* 'lie' take *sein* as their perfect auxiliary, it is categorically ruled out in other varieties, including Western German. In the Western German varieties (just as in the standard language), the only acceptable auxiliary in such contexts is *haben* ('have': *Heute habe ich den ganzen Tag gestanden*). From the perspective of the participants in S1, the construction in (13) is not marginally marked or colloquial, but straightforwardly ungrammatical. Thus, Item 39 presents a particularly stark instance of regional grammatical asymmetry.

As summarized in Table 7 and graphically illustrated in Figure 10, the reaction from S1 speakers confirms this: 61.5% rated the sentence as 'very unnatural', with an additional 20% selecting 'rather unnatural', yielding a mean acceptability score of 1.72:

Table 7. Ratings in S1 for diatopically marked Item 39.

Answer		**Responses**	**Percentage**	**Mean acceptability rating**
1 ('very unnatural')		40	61.54%	1.72
2 ('rather unnatural')		13	20.0%	
3 ('moderately natural')		4	6.15%	
4 ('rather natural')		6	9.23%	
5 ('very natural')		2	3.08%	

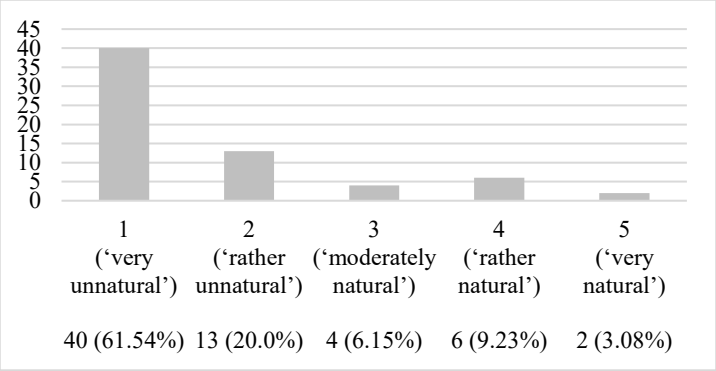

Figure 10. Distribution of S1 ratings for diatopically marked Item 39.

The distribution is sharply skewed toward the lower end of the scale, with only 3.1% of participants selecting the highest naturalness rating. Crucially, ChatGPT-4’s evaluation tracks this distribution closely. It assigned a rating of 2/5 to the same sentence, suggesting sensitivity to the markedness of the auxiliary choice. This convergence is striking: as with the split prepositional adverb in Item 41, the model does not simply assess grammatical well-formedness in abstract structural terms, but appears to reflect the regionalized norms of a specific variety – in this case, Western German.

Items 41 and 39, considered together, suggest a non-trivial diatopic bias in ChatGPT-4’s grammaticality judgments. In both cases, the model produces ratings that mirror the acceptability intuitions of Western German speakers, even though the constructions themselves are attested and well-formed in other regional varieties. The pattern is thus one of selective alignment: structures that are both formally licit and regionally entrenched are not uniformly accepted, but are evaluated in a manner consistent with a particular dialect group. This strongly suggests that ChatGPT-4’s linguistic output is shaped not by a uniform or standard-neutral grammar of German, but by distributional cues embedded in its training data – cues that may disproportionately reflect the norms of dominant or overrepresented varieties. Whether this alignment arises from frequency effects, representational imbalances or corpus composition remains to be explored and cannot be determined on the basis of the present data alone. What is clear at this stage is that the model behaves, in both examples, as if it were a speaker of a western/nothern (or, more broadly, of a *non-southern*) variety of German.

These findings prompt a final and more speculative question: why does ChatGPT-4 align so consistently with the acceptability judgments of Western German speakers, exhibiting tolerance toward regionally marked features from that domain (such as split prepositional adverbs) while sharply rejecting southern features like *sein*-auxiliary selection in the *Perfekt* of posture verbs despite both constructions being attested in naturalistic usage? The explanation cannot simply lie in a preference for standard German, since at least some of the forms treated favourably by the model clearly fall outside codified grammatical norms. Instead,

the model's evaluative behavior appears to reflect asymmetries in how different variants are represented, distributed and framed across the textual corpora on which it was trained. It is plausible that constructions typical of Western and Northern colloquial German occur frequently and unproblematically in informal online writing, including blogs, forums, and social media, where they are embedded in naturally occurring, unmarked usage and thereby receive implicit validation. By contrast, southern variants may be less frequent in such domains and, when they do appear, are more often explicitly flagged as dialectal, non-standard, or erroneous – whether through correction, commentary, or their placement within metalinguistic discourse. The model, in turn, is unlikely to encounter these forms in environments of neutral acceptability. Instead, it is disproportionately exposed to contextual cues that frame them as exceptional or deviant. What thus appears to emerge is not a principled rejection of diatopic variation, but rather a learned evaluative asymmetry: the model seems to replicate not merely the frequency of linguistic forms, but also the attitudinal landscape in which they are embedded. Its internal "grammar" may therefore be shaped not only by usage patterns, but also by what can be described as textual ideology, that is, the ways in which certain variants are made visible, legitimised, or problematized within the digital public sphere. In this light, ChatGPT-4's apparent Western German bias is not simply a matter of geographic alignment, but a product of discursive availability and epistemic framing in the data from which it learns.

6 Conclusions

This study has examined how ChatGPT-4 evaluates a broad spectrum of contextually embedded German utterances, attending not only to grammatical well-formedness in the narrow sense, but also to the wider domain of linguistic acceptability, including pragmatic appropriateness, regional variation and register sensitivity. The findings reveal a notable degree of alignment between the model and human speakers: across all stimulus groups, ChatGPT-4 assigns acceptability ratings that broadly track native intuitions, with no statistically significant divergences at the group level.

At the same time, the qualitative analyses bring to light systematic asymmetries in the model's behavior, particularly in edge cases involving diatopic variation and morphosyntactic markedness. These discrepancies cannot be attributed to misjudgements of structural well-formedness alone, but instead reflect the model's responsiveness – sometimes precise, sometimes skewed – to socially and distributionally mediated norms of acceptability. What becomes visible in the model's output is not a "grammar" in the classical sense, but a gradient evaluative surface shaped by probabilistic exposure, representational imbalances and the discursive framing of linguistic variants in the training data. The model does not merely identify violations of form: it replicates the attitudinal landscape in which certain forms circulate, including patterns of visibility, legitimacy and perceived correctness. In this sense, the study highlights both the potential and the limits of

LLM-based modelling. Such systems can approximate speaker-like judgments, but what they approximate is usage filtered through the social and textual biases of their training environments, not linguistic competence in isolation. Future work will need to explore these biases in greater depth, examining how different corpus compositions, prompting regimes or model architectures may shape the scope and specificity of linguistic evaluation in generative systems.

Conflicts of interest

The author declares no conflicts of interest regarding the publication of this contribution.

References

Tools, dictionaries and corpora

DWDS – Digitales Wörterbuch der deutschen Sprache, ed. by the Berlin-Brandenburgische Akademie der Wissenschaften – entry: 'brustschwimmen'. https://www.duden.de/rechtschreibung/brustschwimmen (last accessed on 12.8.2025).

DWDS – Digitales Wörterbuch der deutschen Sprache, ed. by the Berlin-Brandenburgische Akademie der Wissenschaften – entry: 'kopfstehen'. https://www.dwds.de/wb/kopfstehen (last accessed on 12.8.2025).

DWDS – Digitales Wörterbuch der deutschen Sprache, ed. by the Berlin-Brandenburgische Akademie der Wissenschaften – entry: 'notlanden'. https://www.dwds.de/wb/notlanden (last accessed on 20.8.2025).

DWDS Web Corpus. Text corpus provided by the *Digitales Wörterbuch der deutschen Sprache* (DWDS), https://www.dwds.de/d/korpora/web (last accessed on 1.7.2025).

Duden online, ed. by Dudenredaktion – entry: 'brustschwimmen'. https://www.duden.de/rechtschreibung/brustschwimmen (last accessed on 1.7.2025).

Duden online, ed. by Dudenredaktion – entry: 'notlanden'. https://www.duden.de/rechtschreibung/notlanden (last accessed on 12.8.2025).

Michel, Jean-Baptiste et al. 2011a. *Google Books Ngram Viewer* [Data visualization tool]. Google. https://books.google.com/ngrams (last accessed on 3.8.2025).

JASP Team (2024). JASP (Version 0.95.1)[Computer software]. https://jasp-stats.org/

OpenAI. (2023). ChatGPT (GPT-4) [LLM]. https://chat.openai.com

Survio. Online survey platform. https://www.survio.com.

Secondary literature

Ahlers, Timo. 2010. *Komplexe C°-phobe Verben des Deutschen*. Master's thesis. Universität Wien.

Albrecht, Steffen. 2023. ChatGPT und andere Computermodelle zur Sprachverarbeitung – Grundlagen, Anwendungspotenziale und mögliche Auswirkungen. Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag (TAB). TAB-Hintergrundpapier 26. https://doi.org/10.5445/IR/1000158070

Åsdahl-Holmberg, Märta. 1976. Studien zu den verbalen Pseudokomposita im Deutschen (Götebur ger Germanistische Forschungen 14). Lund: Carl Bloms Boktryckeri.

Austin, Jacob & Odena, Augustus & Nye, Maxwell & Bosma, Maarten & Michalewski, Henryk & Dohan, David & Jiang, Ellen & Cai, Carrie & Terry, Michael & Le, Quoc V. & Sutton, Charles. 2021. Program Synthesis with Large Language Models. *arXiv abs/2108.07732*.

Bang, Yejin & Cahyawijaya, Samuel & Lee, Nayeon & Dai, Wenliang & Su, Dan & Wilie, Bryan & Lovenia, Holy &, Ji, Ziwei & Yu, Tiezheng & Chung, Willy et al. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*. https://doi.org/10.18653/v1/2023.ijcnlp-main.45.

Bavaresco, Anna & Bernardi, Raffaella & Bertolazzi, Leonardo & Elliott, Desmond & Fernández, Raquel & Gatt, Albert et al. 2025. LLMs instead of human judges? A large scale empirical study across 20 NLP evaluation tasks. In Che, Wanxiang & Nabende, Joyce & Shutova, Ekaterina & Taher Pilehvar, Mohammad (eds), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 238–255. Vienna: Association for Computational Linguistics.

Brown, Tom B. & Mann, Benjamin & Ryder, Nick & Subbiah, Melanie & Kaplan, Jared, & Dhariwal, Prafulla et al. 2020. Language Models are few-shot learners. *arXiv preprint arXiv:2005.14165.*

Chen, Lingjiao & Cai, Tracy & Zaharia, Matei & Zou, James. 2021. Did the model change? Efficiently assessing machine learning API shifts. *arXiv preprint arXiv:2107.14203*.

Chen, Lingjiao & Zaharia, Matei & Zou, James. 2023. *How is ChatGPT's behavior changing over time?* Ms., Stanford University and UC Berkeley.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, Noam. 1965. *Knowledge of Language: Its Nature, Origin, and Use*. New York: Praeger.

Cirkel, Philipp & Freywald, Ulrike. 2021. In Stadt und Stadt: Berlin und Ruhrgebiet im Vergleich. *Linguistik Online* 110/5. 193–227. https://doi.org/10.13092/lo.110.8144.

Cohen, L.J. 1981. Some remarks on the nature of linguistic theory. Philosophical Transactions of the Royal Society of London. *Biological Sciences* 295/1077. 235–243. https://doi.org/10.1098/rstb.1981.0136.

De Cesare, Anna-Maria & Bambini, Valentina & Tavosanis, Mirko (eds). 2025. AI-Driven Linguistic Studies: From Text Simplification to Implicit Content Detection, Special Issue in *AI Linguistica* 2/2. https://doi.org/10.62408/ai-ling.v2i2.

Derwing, Bruce L. 1979. Against autonomous linguistics. In Perry, Thomas A. (ed), Evidence and Argumentation in Linguistics, 163–189. Berlin: de Gruyter. https://doi.org/10.1515/9783110848854-010.

Featherston, Sam. 2007. Experimentell erhobene Grammatikalitätsurteile und ihre Bedeutung für die Syntaxtheorie. In Kallmeyer, Werner & Zifonun, Gisela (eds), *Sprachkorpora. Datenmengen und Erkenntnisfortschritt* (= Jahrbuch des Instituts für Deutsche Sprache 2006), 49–69. Berlin: de Gruyter. https://doi.org/10.1515/9783110439083-005.

Fleischer, Jürg. 2002a. *Die Syntax von Pronominaladverbien in den Dialekten des Deutschen: Eine Untersuchung zu Preposition Stranding und verwandten Phänomenen.* Stuttgart: Steiner.

Fleischer, Jürg. 2002b. Preposition stranding in German dialects. In Barbiers, Sief, Cornips, Leonie & van der Kleij, Susanne (eds), *Syntactic Microvariation*, 116–151. Amsterdam: Mertens Instituut.

Forche, Christian. 2020. *NonV2-Verben im Deutschen: Theoretische Überlegungen und empirische Untersuchungen zu einem morphosyntaktischen Problemfall (den es vielleicht gar nicht gibt)*. Berlin & Heidelberg: Springer. https://doi.org/10.1007/978-3-662-61926-1.

Fortmann, Christian. 2007. Bewegungsresistente Verben. *Zeitschrift für Sprachwissenschaft* 26. 1–40. https://doi.org/10.1515/ZFS.2007.009.

Fortmann, Christian. 2015. Verbal pseudo-compounds in German. In Müller, Peter O. & Ohnheiser, Ingeborg & Olsen, Susan & Rainer, Franz (eds.), *Word-Formation. An International Handbook of the Languages of Europe*, Vol. 1, 594–610. Berlin: de Gruyter. https://doi.org/10.1515/9783110246254-036.

Freywald, Ulrike & Simon, Horst. 2007. Wenn die Wortbildung die Syntax stört: Über Verben, die nicht in V2 stehen können. In Kauffer, Maurice & Métrich, René (eds), *Verbale Wortbildung im Spannungsfeld zwischen Wortsemantik, Syntax und Rechtschreibung*, 181–194. Tübingen: Stauffenburg.

Gibson, Edward & Thomas, James. 1999. Memory limitations and structural forgetting: the perception of complex ungrammatical sentences as

grammatical. *Language and Cognitive Processes* 14. 225–248. https://doi.org/10.1080/016909699386293.

Grice, H. Paul. 1975. *Logic and conversation*. In Cole, Peter & Morgan, Jerry (eds), *Syntax and Semantics*, 41–58. New York: Academic Press. https://doi.org/10.1163/9789004368811_003.

Gross, Steven. 2021. Linguistic judgments as evidence. In Allott, Nicholas & Lohndal, Terje & Rey, Georges (eds), *Blackwell Companion to Chomsky*, 544–556. Hoboken, NJ: Wiley-Blackwell. https://doi.org/10.1002/9781119598732.ch3.

Hu, Jennifer & Mahowald, Kyle & Lupyan, Gary & Ivanova, Anna & Levy, Roger. 2024. Language models align with human judgments on key grammatical constructions. *Psychological and Cognitive Sciences* 121/36. 1–3. https://doi.org/10.1073/pnas.2400917121

Ide, Yusuke & Nishida, Yuto & Vasselli, Justin & Oba, Miyu & Sakai, Yusuke & Kamigaito, Hidetaka & Watanabe, Taro. 2025. How to make the most of LLMs' grammatical knowledge for acceptability judgments. In Chiruzzo, Luis & Ritter, Alan & Wang, Lu (eds), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, 7416–7432.

Johnsen, Lars G. B. (2025). Grammaticality judgments in humans and language models: Revisiting Generative Grammar with LLMs. https://arxiv.org/abs/2512.10453

Kahnemann, Daniel. 2011. *Thinking, fast and slow*. London: Penguin Books.

Lau, Jey Han & Clark, Alexander & Lappin, Shalom. 2016. Grammaticality, acceptability, and probability: a probabilistic view of linguistic knowledge. *Cognitive Science* 41. https://doi.org/10.1111/cogs.12414.

Leivada, Evelina & Westergaard, Marit. 2020. Acceptable ungrammatical sentences, unacceptable grammatical sentences, and the role of the cognitive parser. *Frontiers in Psychology* 11/364. 1–9. https://doi.org/10.3389/fpsyg.2020.00364

Leivada, Evelina & Günther, Fritz & Dentella, Vittoria. 2024. Reply to Hu et al.: Applying different evaluation standards to humans vs. Large Language Models overestimates AI performance. *Psychological and Cognitive Sciences* 121/36. 1–2. https://doi.org/10.1073/pnas.2406752121

Leser, Stephanie. 2012. Zum Pronominaladverb in den hessischen Dialekten. Eine Untersuchung zum Verlauf syntaktischer Isoglossen. In Langhanke, Robert & Berg, Kristian & Elmentaler, Michael & Peters, Jörg (eds), *Niederdeutsche Syntax. Germanistische Linguistik*, 79–99. Hildesheim: Olms.

Levelt, Willem J. M. & van Gent, J. A. W. M. & Haans, A. F. J. & Meijers, A. J. A. 1977. Grammaticality, paraphrase, and imagery. In Greenbaum, Sidney

(ed), *Acceptability in Language*, 87–101. The Hague: Mouton. https://doi.org/10.1515/9783110806656-008.

Li, Junyi & Tang, Tianyi & Zhao, Wayne Xin & Nie, Jian-Yun & Wen, Ji-Rong. 2022. Pretrained language models for text generation: A survey. *arXiv:2201.05273v4*. https://doi.org/10.48550/arXiv.2201.05273.

Liu, Hanmeng & Ning, Ruoxi & Teng, Zhiyang & Liu, Jian & Zhou, Qiji & Zhang, Yue. 2023. Evaluating the logical reasoning ability of ChatGPT and GPT-4. *arXiv preprint arXiv:2304.03439*.

Marvin, Rebecca & Linzen, Tal. 2018. Targeted syntactic evaluation of language models. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics

Masood, Faiza & Khan, Farzana. 2025. Assessing linguistic naturalness in ChatGPT-generated English: A comparative study with human written texts. *Qualitative Research Journal for Social Studies* 2/2. 1046–1056. https://qrjsocial.com/index.php/38/article/view/145/133 (last accessed on 1.8.2025).

Meier, Franz. 2024. Dealing with common ground in Human Translation and Neural Machine Translation: A case study on Italian equivalents of German modal particles. AI Linguistica 1/1. 1–20. https://doi.org/10.62408/ai-ling.v1i1.12.

Meinunger, André. 2022. Pre-field phobia – About formal and meaning-related prohibitions on starting a German V2 clause. *The Linguistic Review* 39/4. 699–742. https://doi.org/10.1515/tlr-2022-2102.

Michel, Jean-Baptiste et al. 2011b. Quantitative analysis of culture using millions of digitized books. *Science* 331. 176–182. https://doi.org/10.1126/science.1199644.

Negele, Michaela. 2012. *Varianten der Pronominaladverbien im Neuhochdeutschen: Grammatische und soziolinguistische Untersuchungen*. Berlin: de Gruyter. https://doi.org/10.1515/9783110273281.

Peters, Matthew E. & Neumann, Mark & Iyyer, Mohit & Gardner, Matt & Clark, Christopher, & Lee, Kenton et al. 2018. Deep Contextualized Word Representations. In Walker, Marilyn & Ji, Heng & Stent, Amanda (eds), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Vol. 1, 2227–2237. New Orleans: Association for Computational Linguistics.

Qiu, Zhuang & Duan, Xufeng & Cai, Zhenguang. 2024. Evaluating grammatical well-formedness in large language models: A comparative study with human judgments. In Kuribayashi, Tatsuki & Rambelli, Giulia & Takmaz, Ece & Wicke, Philipp & Oseki, Yohei (eds), Proceedings of the Workshop on

Cognitive Modeling and Computational Linguistics, 189–198. Bangkok: Association for Computational Linguistics.

Radford, Alec & Narasimhan, Karthik & Salimans, Tim & Sutskever, Ilya. 2019. Improving language understanding by generative pre-training. *OpenAI*. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (last accessed on 19.8.2025).

Schütze, Carson T. 2016. The Empirical Base of Linguistics. Grammaticality Judgments and Linguistic Methodology (= Classics in Linguistics 2). Berlin: Language Science Press. https://doi.org/10.26530/OAPEN_603356.

Sperber, Dan & Wilson, Deirdre. 1998. The mapping between the mental and the public lexicon. In Carruthers, Peter & Boucher, Jill (eds), *Thought and Language*, 184–200. Cambridge: Cambridge University Press.

Sprouse, Jon. 2018. Acceptability judgments and grammaticality, prospects and challenges. In Hornstein, Norbert & Lasnik, Howard & Patel-Grosz, Pritty & Yang, Charles (eds), Syntactic *Structures After 60 Years. The Impact of the Chomskyan Revolution in Linguistics* (= Studies in Generative Grammar 129), 195–223. Berlin: de Gruyter. https://doi.org/10.1515/9781501506925-199

Sternefeld, Wolfgang. 2008. Syntax. Eine morphologisch motivierte generative Beschreibung des Deutschen, Vol. 1. Tübingen: Stauffenburg.

Tikhonova, Elena & Raitskaya, Lilia. 2023. Exploring the realm of GPT and large language models. *Journal of Language and Education* 9/3. 5–11. https://doi.org/10.17323/jle.2023.18119.

van Dijk, Teun A. 1977. *Text and Context: Explorations in the Semantics and Pragmatics of Discourse*, London & New York: Longman.

Vikner, Sten. 2005 Immobile complex verbs in Germanic. *Journal of Comparative Germanic Linguistics* 8/1-2. 83–115. https://doi.org/10.1007/s10828-004-0726-9.

Wellwood, Alexis & Pancheva, Roumyana & Hacquard, Valentine & Phillips, Colin 2018. The anatomy of a comparative illusion. *Journal of Semantics* 35. 543–583. https://doi.org/10.1093/jos/ffy014.

Yu, Danni & Li, Luyang & Su, Hang & Fuoli, Matteo. im Ersch. Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apologies. *International Journal of Corpus Linguistics* 29/4. 534–561. https://doi.org/10.1075/ijcl.23087.yu.

Zhou, Houquan & Hou, Yang & Li, Zhenghua & Wang, Xuebin & Wang, Zhefeng & Duan, Xinyu & Zhang, Min. 2023. How well do Large Language Models understand syntax? An evaluation by asking natural language questions. https://doi.org/10.48550/arXiv.2311.08287.