

Article

Norms, Contexts and Patterns of Variation: Evaluating Acceptability Judgments of Five LLMs Across Linguistic Dimensions in German

Nicholas Catasso *  and Finn Esser

Faculty of Humanities and Cultural Studies, Institute of Linguistics/German Studies,
Bergische Universität Wuppertal, 42119 Wuppertal, Germany; finn.esser@uni-wuppertal.de

* Correspondence: catasso@uni-wuppertal.de

Abstract

This paper reports on a pilot study evaluating five large language models (ChatGPT-4, Gemini 2.0 Flash, Claude 3.5 Sonnet, Perplexity AI, and DeepSeek) in gradient acceptability judgment tasks in German. The models rated 150 contextually embedded sentences on a 5-point Likert scale across five categories: gray-zone (variable) items, canonical grammatical items, ungrammatical items, diatopically marked items, and diastratically/diaphasically marked items. All models clearly distinguish between clearly grammatical and clearly ungrammatical stimuli in unambiguous morphosyntactic contexts. Mixed-effects analyses further show that differences between models vary across stimulus categories rather than reflecting a uniform global shift in acceptability ratings. These findings indicate that current LLMs robustly capture core morphosyntactic contrasts, but that model behavior is less uniform in domains involving variation and contextual sensitivity. The study contributes to the empirical assessment of LLMs as acceptability raters and informs debates on their methodological role in linguistics.

Keywords: large language models; acceptability judgments; grammaticality; variation; sociolinguistics; German; model evaluation

1. Introduction

The recent evolution of Artificial Intelligence (henceforth: AI) has brought about a paradigm shift in how language is processed, modeled and studied. At the forefront of this development are large language models (henceforth: LLMs) such as ChatGPT, Gemini, Claude, and others, which rely on massive neural architectures trained on vast corpora of text data. These models have demonstrated remarkable fluency, syntactic competence, and contextual awareness, sparking considerable interest in their capacity to perform tasks traditionally associated with human linguistic cognition. One such task is the assessment of grammaticality and acceptability, which lies at the heart of both theoretical and empirical linguistics.

Acceptability judgments are a key empirical tool in syntactic theory, providing insight into speakers' internalized grammar and their intuitions about linguistic well-formedness. In traditional formalist frameworks, grammaticality has often been treated as a binary property of linguistic form: a sentence either conforms to the rules of grammar or it does not [1–3]. However, decades of research have shown that speakers' judgments are far more nuanced. Acceptability is influenced not only by morphosyntactic structure,



Academic Editor: Juraj Gregaň

Received: 27 January 2026

Revised: 6 March 2026

Accepted: 8 March 2026

Published: 16 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

but also by contextual appropriateness, pragmatic felicity, discourse embedding, register, and sociolinguistic variation [4–8]. As a result, linguistic well-formedness is now widely recognized as gradient rather than categorical, and tightly linked to interface phenomena across syntax, semantics, and pragmatics.

In parallel, LLMs have become powerful tools for modeling such judgments. Prior research has shown that models like GPT-3 and GPT-4 can often assign plausibly human-like ratings to a range of grammatical constructions [9–14], and that their evaluations frequently correlate with those of native speakers—especially for prototypically grammatical or ungrammatical sentences. However, these studies have typically focused on English, isolated individual models, and have avoided constructions that involve variation, sociolinguistic markedness, or context-dependent meaning. As a result, current understanding of how LLMs perform on more nuanced acceptability tasks—particularly in languages other than English and domains shaped by social and regional norms—remains limited.

The present study aims to address these gaps through a systematic cross-model evaluation of five LLMs—ChatGPT-4, Gemini 2.0 Flash, Claude 3.5 Sonnet, Perplexity AI and DeepSeek—on a gradient acceptability judgment task in German. Unlike previous research, which has largely focused on categorical contrasts or grammaticality in isolation, this study adopts a sociolinguistically enriched and pragmatically contextualized design, evaluating the models across a range of morphosyntactic, diatopic (regional), diastratic/diaphasic (social/register-related), and variation-sensitive constructions. Each model was prompted to rate a set of 150 contextually embedded sentences on a 5-point Likert scale, allowing for fine-grained scalar assessments. The test items were carefully constructed and distributed across five linguistic domains: (1) canonical grammaticality, (2) clear ungrammaticality, (3) standard forms with documented variation, (4) regionally marked structures, and (5) socially or stylistically marked expressions.

The choice of German as a target language is methodologically significant. German offers a rich array of morphosyntactic contrasts, as well as a documented landscape of dialectal, sociolectal and register-based variation. From verb-second (V2) phenomena and case assignment to regional adverb splitting and register-sensitive vocabulary, German presents a robust testing ground for linguistic evaluation. Importantly, the language exhibits sociolinguistic asymmetries that can serve as sensitive indicators of a model's exposure to and handling of variation. Certain constructions, for instance, are perfectly acceptable in colloquial usage or in regional varieties but are penalized in normative contexts or by speakers of standard varieties. Whether LLMs can recognize and reflect these differences offers crucial insight into the scope and bias of their internal representations.

Beyond its focus on language-specific phenomena, the present study also seeks to contribute to broader debates about the epistemological and methodological role of LLMs in linguistic research. There is growing interest in whether such models can serve not only as tools for automation (e.g., in annotation or corpus expansion) but also as theoretical probes into human linguistic competence. Can LLMs serve as reliable proxies for speaker intuitions? Do they internalize grammatical structure or merely reproduce surface patterns? Are they sensitive to the gradient, probabilistic nature of acceptability—or do they default to categorical or prescriptive norms? These questions have immediate consequences for how LLMs are used in linguistics, particularly in experimental settings where speaker data may be costly or difficult to obtain.

In adopting a comparative, scalar, and variation-sensitive approach, this study not only broadens the empirical base for evaluating LLMs in linguistic tasks but also advances our understanding of the limits and affordances of AI-based models in capturing the richness of natural language as it is structured, used, and socially embedded.

The remainder of this article is organized as follows. Section 2 describes the study design and data collection in detail: it introduces the five evaluated LLMs and their access conditions, specifies the German prompting protocol and session controls and documents the construction of the 150 contextually embedded stimuli across the five item types (G1–G5), as well as the logging, parsing and preprocessing steps applied to the model outputs. Section 3 reports the quantitative analyses of the resulting 750 ratings, presenting the omnibus test for overall model differences and the post hoc pairwise comparisons used to address RQ1 and RQ2. Section 4 interprets these results with a focus on what aggregated null effects do (and do not) imply and situates the findings in current work on LLMs as acceptability raters, especially with respect to normativity, register and sociolinguistic variation. Section 5 summarizes the main contributions, acknowledges limitations of the pilot design and outlines concrete next steps.

2. Materials and Methods

2.1. Overview and Research Questions

The aim of this study is to evaluate the behavior of current LLMs when performing gradient acceptability judgments in German. Specifically, the study compares the responses of five widely available LLMs to a set of 150 sentences, each embedded in a brief discourse context and covering a range of grammatical, ungrammatical, variation-sensitive and sociolinguistically marked constructions. The experimental design follows a scalar judgment format, in which the models were instructed to rate each sentence on a five-point Likert scale, ranging from completely unacceptable (1) to completely acceptable (5).

This study was motivated by the need to assess not only whether LLMs can distinguish between grammatical and ungrammatical input but also whether they are sensitive to gradient, register variation, diatopic and diastratic markers, and other linguistic factors that shape human acceptability judgments. In contrast to much of the existing work—often limited to a single model, a narrow syntactic domain, or a single language variety—this study adopts a variation-aware approach, with particular emphasis on sociolinguistic diversity in German.

In particular, the following two research questions constitute the central objectives of the present study:

- RQ1: Do the evaluated LLMs differ in their overall acceptability rating behavior when ratings are aggregated across all stimulus types (i.e., is there an overall model effect on mean acceptability ratings)?
- RQ2: Are there statistically significant differences in acceptability judgments between individual LLMs (e.g., ChatGPT vs. Gemini, ChatGPT vs. Claude, ChatGPT vs. Perplexity, Gemini vs. Claude)?

All five models were evaluated independently and under consistent input conditions. The full set of stimuli, prompt formats, and model outputs is included in the Supplementary Materials and is also made available via a public open-access repository. In the interest of transparency and reproducibility, all scripts used for querying, formatting, and analyzing the data are also provided.

This study evaluates cross-model internal consistency and bias patterns in LLM acceptability ratings. It does not aim to establish equivalence to human judgments, which is addressed in separate work using a complementary dataset (e.g., [15]).

The methodological procedure is structured as follows. Section 2.2 describes the language models examined in this study, including their access modalities and deployment conditions. Section 2.3 outlines the prompting procedure used to elicit scalar acceptability judgments from the models. Section 2.4 details the design of the stimulus set, including the linguistic phenomena targeted and the categorization of items according to grammatical, diatopic and diastratic dimensions of variation. Finally, Section 2.5 documents the data col-

lection process, including session management, response logging, and basic preprocessing steps applied prior to analysis.

2.2. Language Models

This study examined the behavior of five state-of-the-art LLMs, selected for their accessibility, performance, and architectural diversity. All models were publicly available at the time of data collection and could be accessed either via a web interface or an application programming interface (API). The selection criteria aimed to reflect a representative cross-section of leading LLMs with multilingual capabilities, with a focus on those likely to have been trained on German or German-adjacent corpora.

The models included in this study are: ChatGPT-4 [16], Gemini 2.0 Flash [17], Claude 3.5 Sonnet [18], Perplexity AI [19] and DeepSeek Chat [20]. Each model was queried individually using prompt templates developed and standardized within the study, thereby ensuring consistent instructions across systems. Where necessary (that is, if multiple), care was taken to reset the session history between queries to avoid cross-item contamination or memory effects.

ChatGPT-4, accessed via the OpenAI ChatGPT platform (<https://chat.openai.com>, accessed on 27 January 2026), was used in its GPT-4-turbo configuration, which was the default deployment at the time of testing in 2025. All inputs were submitted via the chat interface. Evaluations were conducted using the standard deployment settings provided by the respective web interfaces. No manual adjustments to temperature or sampling parameters were made. Gemini 2.0 Flash was accessed via the Google Gemini platform (<https://gemini.google.com>, accessed on 27 January 2026) in its “Flash” variant, which prioritizes faster response times. While Gemini’s architecture is not fully disclosed, it is presumed to incorporate transformer-based design principles similar to other LLMs in this study. As with ChatGPT, prompts were submitted manually via the browser interface. Claude 3.5 Sonnet was accessed via Anthropic’s Claude platform (<https://claude.ai>, accessed on 27 January 2026). This model represents a mid-sized configuration in the Claude 3 family and is designed for general-purpose reasoning and language understanding. Prompts were entered through the web interface, and session history was cleared between batches of inputs. Perplexity AI was accessed at <https://www.perplexity.ai>, accessed on 27 January 2026 using the default free-tier configuration. The platform functions as a multi-model interface and may dynamically route queries to different underlying models (e.g., GPT-4, Claude, Mixtral). In this study, web search capabilities were disabled to ensure that the models relied solely on their internal representations rather than retrieving information online. Note that explicit locking to a single underlying model architecture was not available in the interface used for data collection, so it cannot be excluded that individual queries were dynamically routed to different backend models. Perplexity is therefore treated in the present study as a deployed LLM-based platform rather than as a uniquely identifiable model instance. The results are interpreted as reflecting the behavior of the publicly accessible system at the time of testing. The objective of this comparison is to evaluate publicly accessible LLM systems under real-world access conditions rather than to isolate fixed underlying model checkpoints. Accordingly, any internal routing variability is interpreted as part of the observable system behavior during the data-collection window rather than as an external methodological confound. DeepSeek Chat was accessed via <https://chat.deepseek.com>, accessed on 27 January 2026. While less documented than the other models, DeepSeek represents an emerging architecture trained on multilingual corpora and optimized for performance in multiple languages. Its inclusion was motivated by the need to explore variation across less Western-centric model designs. All models were instructed using equivalent language and structure, translated and adapted for German where appropriate. Model responses were manually recorded and stored in structured formats for analysis. Cf. Table 1:

Table 1. Overview of publicly accessible LLM systems evaluated in this study, including deployment interfaces and provider-documented architectural information.

Provider	System/Deployment Name	Access Modality	Architecture Class (Public Documentation)
OpenAI	ChatGPT-4	Web chat UI	Transformer-based autoregressive LLM
Google	Gemini 2.0 Flash	Web chat UI	Transformer-based LLM
Anthropic	Claude 3.5 Sonnet	Web chat UI	Transformer-based LLM
Perplexity	Perplexity AI, default model (Sonar)	Web chat UI	Transformer-based LLM system
DeepSeek	DeepSeek (DeepSeek-V3)	Web chat UI	Transformer-based LLM

Note that architecture class assignments are based on publicly available provider documentation and refer to model families or deployed systems rather than uniquely identifiable checkpoints. The evaluated systems were accessed via proprietary web-based interfaces, which do not expose stable checkpoint identifiers, parameter counts, training data cutoffs, or fine-grained architectural specifications. Therefore, the present study reports only verifiable deployment-level information.

2.3. Prompting Procedure

Each LLM was prompted using a standardized instruction in German, asking the model to evaluate each sentence within its given context using a five-point Likert scale. The instruction emphasized that judgments should reflect naturalness and linguistic appropriateness rather than strict grammatical correctness.

In the study presented here, the models were prompted using a single, unified prompt that included both the general instructions and the complete set of sentences to be rated. This approach was chosen to maintain consistency and minimize variation in system behavior across items. As an exploratory robustness check, a subset of 50 items (10 per stimulus category) was also submitted using separate prompts for each sentence. This comparison was not designed as a fully systematic positional-effects analysis, but rather as a preliminary check for gross deviations between batched and single-item prompting. No systematic differences in mean rating patterns were observed. However, a formal analysis of potential recency or drift effects in long-context prompting was not conducted and remains an important direction for future work.

The introductory section of this prompt, which provided the evaluation guidelines, was as follows (see English translation below):

Im Folgenden wird Ihnen eine Reihe von Sätzen auf Deutsch präsentiert. Ihre Aufgabe besteht darin, zu beurteilen, wie natürlich und korrekt diese Sätze für Sie klingen.

Jeder Satz soll auf einer Skala von 1 bis 5 bewertet werden. Eine Bewertung von 1 bedeutet, dass der Satz sehr unnatürlich oder falsch klingt—zum Beispiel so, als würde er im alltäglichen Sprachgebrauch normalerweise nicht verwendet oder als wäre er grammatikalisch fehlerhaft. Eine Bewertung von 5 bedeutet, dass der Satz völlig natürlich und korrekt klingt, wie man es im standardsprachlichen Alltagsgebrauch erwarten würde.

Die Zwischenwerte 2, 3 und 4 stehen für unterschiedliche Abstufungen von Natürlichkeit und Korrektheit: 2 deutet darauf hin, dass der Satz eher unnatürlich oder fehlerhaft klingt, 3 spiegelt ein mittleres Maß an Akzeptabilität wider, und 4 zeigt an, dass der Satz eher natürlich oder korrekt klingt.

‘In the following, you will be presented with a series of sentences in German. Your task is to assess how natural and correct these sentences sound to you.

Each sentence should be rated on a scale from 1 to 5. A rating of 1 indicates that the sentence sounds very unnatural or incorrect—for example, as though it would not typically be used in everyday language or would appear grammatically wrong. A rating of 5 indicates that the sentence sounds entirely natural and correct, as one would expect in standard, everyday usage. The intermediate values 2, 3, and 4 represent varying degrees of naturalness and correctness: 2 suggests the sentence sounds rather unnatural or incorrect, 3 reflects a moderate degree of acceptability, and 4 indicates that the sentence sounds rather natural or correct.’

2.4. Stimulus Design

The stimulus set consisted of 150 sentence items, each embedded within a minimal discourse context of one to two sentences. This design aimed to simulate naturalistic language use while minimizing interpretive ambiguity. All stimuli were constructed by the author, drawing on descriptive linguistic literature, grammatical reference works and known areas of sociolinguistic or morphosyntactic variation in German.

The items were evenly distributed across five target categories, with 30 utterances per category:

- utterances containing structures and forms known to be variable in acceptability among native speakers (so-called “gray-zone items”, German *Zweifelsfälle*; cf., e.g., [21,22]);
- grammatically well-formed (grammatical) utterances;
- clearly ill-formed (ungrammatical) constructions (violating word order, agreement or selection);
- grammatically acceptable utterances containing regionally (diatopically) marked structures;
- grammatically acceptable utterances containing socially / stylistically (diastratically) marked structures and forms.

Each sentence was accompanied by a brief context that framed the communicative situation, allowing the LLMs—just as would be expected with human participants—to interpret the utterance within a meaningful discourse environment. This was particularly important for reducing ambiguity and enhancing ecological validity, especially for sentences targeting phenomena of language variation, where appropriate contextualization was essential to ensure that the model’s responses reflected sociolinguistically informed interpretation rather than decontextualized grammatical judgment. All items were designed to elicit gradient acceptability ratings rather than categorical judgments. The five target categories are outlined in Table 2 and are individually exemplified in the following sections:

Table 2. Categories of stimuli provided to the LLMs.

Group	Label	Definition	Description
G1	Gray-zone items	Grammatical constructions subject to variable acceptability among native speakers.	Productively used forms that often trigger inconsistent intuitions due to norm conflicts, optionality, register variation, or stylistic markedness.
G2	Grammatical items	Sentences that are unambiguously grammatical and pragmatically neutral.	Canonical constructions conforming to standard morphosyntactic norms, broadly accepted across dialects, registers, and discourse contexts.
G3	Ungrammatical items	Sentences that contain clear violations of morphosyntactic rules.	Includes categorical errors involving word order, agreement, case marking, or other core syntactic features of Standard German.

Table 2. Cont.

Group	Label	Definition	Description
G4	Diatopically marked items	Sentences featuring regionally specific lexical or grammatical elements.	Grammatical within certain dialects or regional varieties but often perceived as non-standard or marked by speakers outside those regions.
G5	Diastratically/diaphasically marked items	Sentences characterized by socially or stylistically distinctive features.	Associated with particular social groups or informal registers (e.g., youth language, colloquial speech), and potentially deemed inappropriate in formal contexts.

2.4.1. Gray-Zone Items (G1)

This category includes constructions whose grammatical status is subject to both intra-speaker and inter-speaker variability, often reflecting ongoing processes of morphosyntactic reanalysis. The stimulus items in this group target verbs that exhibit structural ambiguity with respect to separability, a phenomenon that is particularly prominent in compound or recently lexicalized verb forms (cf., e.g., [23–25]). Examples (1)–(2) illustrate minimal pairs (all included in the dataset) involving the verb *bauchreden* (‘to ventriloquize’, lit. ‘stomach-talk’, cf. [26,27]), embedded in a uniform discourse context.

As shown in the examples, this verb is used by some speakers such that the first compound member *bauch-*(cf. *Bauch* ‘stomach’) is syntactically separable (1a), while others treat it as inseparable from the verbal stem (1b). Similarly, the past participle may be formed either as in (2a), with a circumfix (*ge-...-et*; glossed as PP) surrounding only the verbal stem *red-*(‘talk’), or as in (2b), surrounding the entire compound formation. Some speakers do not accept either of the options illustrated in (1) and (2), respectively, while others accept all of them.

For reasons of space, the context is provided in English translation only here and in the following paragraphs:

- (1) Context: One of your friends refers to a person known for using an unusual or rare form of communication:
- a. *Er redet bauch – das ist ja schon*
 he talks stomach that is PRT PRT
was Besonderes!
 something special
- b. *Er bauchredet – das ist ja schon was*
 he stomach-talks that is PRT PRT something
Besonderes!
 special
 ‘He ventriloquizes—that is definitely something special!’
- (2) Context: One of your friends refers to a person known for using an unusual or rare form of communication:
- a. *Hat er bauchgeredet oder habe ich*
 has he stomach-PP.talk.PP or have I
mich verhört?
 REFL misheard
- b. *Hat er gebauchredet oder habe ich*
 has he PP.stomach-talk.PP or have I
mich verhört?
 REFL misheard
 ‘Did he ventriloquize or did I mishear?’

To document the structural properties of the G1 stimulus set, sentence length and lexical diversity were calculated. Across the 30 items in this category, mean sentence length was 8.50 tokens (SD = 2.96, range = 4–14). The set comprised 255 tokens in total, of which 121 were unique types (type–token ratio = 0.48). These values indicate moderate lexical diversity and relatively compact sentence structure within this condition.

2.4.2. Grammatical Items (G2)

Items in this condition were designed to serve as structural and pragmatic baselines, exemplifying uncontroversially grammatical, contextually neutral usage. These sentences reflect the core norms of Standard German syntax and morphology and are expected to yield uniformly high acceptability across speaker groups and models. Example (3) illustrates such an item:

- (3) Context: A fashion designer presents her new collection. She says:

<i>Die Farben</i>	<i>und</i>	<i>Stoffe</i>	<i>dieser</i>	<i>Saison</i>	<i>wurden</i>
the colors	and	fabrics	of-this	season	were
<i>gezielt</i>	<i>ausgewählt,</i>	<i>um einen</i>	<i>eleganten,</i>	<i>aber gleichzeitig</i>	
deliberately	chosen	for an	elegant	but at-the-same-time	
<i>modernen Look</i>	<i>zu kreieren.</i>				
modern look	to create				

‘The colors and fabrics of this season were deliberately chosen to create a look that is both elegant and modern.’

The sentence is well-formed at all levels: morphosyntactically, it adheres to prescriptive norms; pragmatically, it fits a plausible discourse context; and stylistically, it maintains register-appropriate formulation. Items in this group function as controls for detecting deviance in other conditions and verifying the model’s sensitivity to standard constructions.

The 30 grammatical baseline items (G2) had a mean sentence length of 13.17 tokens (SD = 4.40, range = 7–22). The set comprised 395 tokens in total, of which 263 were unique types (type–token ratio = 0.67). Compared to G1, grammatical baseline items were structurally longer and lexically more diverse.

2.4.3. Ungrammatical Items (G3)

Stimuli in this group were explicitly constructed to violate core morphosyntactic rules of Standard German. The aim was to isolate the grammatical dimension of acceptability by embedding each ungrammatical sentence in a coherent context free from discourse-level anomalies. Example (4) demonstrates a case involving multiple morphosyntactic violations:

- (4) Context: A biologist is explaining insect reproduction. He says:

<i>Die Insekten legt</i>	<i>ihre Eier</i>	<i>unter die Baum</i>	<i>und</i>
the insects lays	their eggs	under the tree	and
<i>fliegt weg.</i>			
fly away			

(intended:) ‘The insect(s) lay(s) its/their eggs under the tree and fly/flyes away.’

Two critical errors are present: (1) Subject–verb agreement is violated (*Die Insekten legt* ‘the insects lay.SG’ instead of *legen* ‘lay.PL’), and (2) a gender mismatch occurs (*unter die Baum* ‘under the.F.ACC tree’ instead of *unter den Baum* ‘under the.M.ACC tree’).

These violations collectively render the sentence categorically ungrammatical and are expected to elicit low ratings from both human and model evaluators. The G3 items thus serve to verify whether the model appropriately identifies and penalizes syntactic ill-formedness, independent of contextual or pragmatic felicity.

The 30 ungrammatical items (G3) had a mean sentence length of 11.40 tokens (SD = 1.73, range = 9–16) and a type–token ratio of 0.65 (342 tokens; 223 unique types).

2.4.4. Diatopically Marked Items (G4)

This category includes constructions that are fully grammatical within regional varieties of German but may be perceived as non-standard or marked by speakers from outside those dialect areas. The test items were drawn from colloquial regional German, rather than dialect proper, to preserve accessibility for language models while still retaining salient regional features. Example (5) illustrates such an instance:

- (5) Context: Your friend is talking about a situation that really does not bother him at all.

<i>Da</i>	<i>hab</i>	<i>ich</i>	<i>überhaupt</i>	<i>kein</i>	<i>Problem</i>	<i>mit!</i>
there	have	I	at-all	no	problem	with

‘I’ve got absolutely no problem with that!’

The relevant construction in (5) concerns the realization of the prepositional adverb *damit* (‘with that’) in the predicate expression *damit kein Problem haben* (‘to have no problem with that’). In Standard German, this form typically appears as a contiguous unit within the clause, for example: *Ich habe damit kein Problem* or *Damit habe ich kein Problem*. In the example under discussion, the prepositional adverb appears in a split configuration: the deictic element *da* is fronted (*Da hab ich. . .*), while the preposition *mit* surfaces clause-finally (*. . .kein Problem mit!*). This discontinuous placement reflects a feature characteristic of colloquial Western and Northern varieties of German, where prepositional adverbs such as *damit* (‘with that’), *dafür* (‘for that’), *davon* (‘of that’), or *dagegen* (‘against that’) may be optionally split in informal spoken registers. In many other regional varieties and dialects, however, such splitting is either dispreferred or only licensed under more constrained syntactic or prosodic conditions [28–32]. Morphologically, these forms consist of a locative or referential *da*-element and a governing preposition, whose adjacency is typically preserved in the standard variety. The example thus illustrates regionally marked syntax that remains grammatical and pragmatically appropriate in certain dialectal contexts but may be judged as non-standard or degraded by speakers outside those varieties. Items of this kind are used to assess whether the model can recognize such diatopic variation and appropriately modulate its acceptability judgments based on regional usage patterns.

The 30 diatopically marked items (G4) had a mean sentence length of 11.23 tokens (SD = 3.87, range = 4–24) and a type–token ratio of 0.72 (337 tokens; 244 unique types).

2.4.5. Diastratically/Diaphasically Marked Items (G5)

The final group targets constructions that are grammatically well-formed but strongly marked with respect to social register, group identity, or stylistic tone. These items were designed to probe evaluators’ awareness of pragmatic appropriateness, especially in formal or institutional discourse settings. Example (6) represents a prototypical case of register mismatch:

Context: Two 14-year-old boys are behind the gym smoking a joint, spitting and

- (6) throwing dumb lines at each other. After a while, one of them starts talking about a meetup with his friends:

<i>Gestern</i>	<i>Abend</i>	<i>begaben</i>	<i>wir</i>	<i>uns</i>	<i>zu</i>	<i>einer</i>
yesterday	evening	convened	we	REFL	to	an
<i>ungezwungenen</i>	<i>Zusammenkunft,</i>	<i>um</i>	<i>in</i>	<i>gelöster</i>	<i>Atmosphäre</i>	
informal	gathering	for	in	congenial	atmosphere	
<i>mannigfaltige</i>	<i>Aspekte des</i>	<i>Alltagsgeschehens</i>	<i>zu</i>	<i>erörten.</i>		
various	facets	of-the	quotidian-affairs	to	deliberate	

‘Yesterday evening, we partook in an informal gathering to deliberate upon various facets of quotidian affairs in a congenial and relaxed atmosphere.’

The sentence, though grammatically flawless, is stylistically incongruent with the preceding context. Elevated lexical choices (e.g., *sich begeben*, *ungezwungene Zusammenkunft*, *mannigfaltige Aspekte*, *Alltagsgeschehen*, *erörtern*) clash with the adolescent, street-level setting,

creating a deliberate register mismatch. These expressions are intended to elicit low acceptability ratings not due to syntactic errors but because of their marked social and stylistic inappropriateness. Of the 30 items in this group, 15 were context-congruent and 15 (like (6)) deliberately violated the contextual norms.

The 30 diastatically and diaphasically marked items (G5) had a mean sentence length of 14.53 tokens ($SD = 7.02$, range = 6–44) and a type–token ratio of 0.63 (436 tokens; 273 unique types).

2.4.6. Descriptive Statistics of Stimulus Properties

For ease of comparison across conditions, Table 3 summarizes the descriptive statistics reported in Sections 2.4.1–2.4.5. The table provides sentence length (mean, SD , range) and lexical diversity (type–token ratio) for each stimulus category:

Table 3. Descriptive statistics of stimulus properties by category.

Group	N	Mean Length	SD	Range	TTR
G1	30	8.50	2.96	4–14	0.48
G2	30	13.17	4.40	7–22	0.67
G3	30	11.40	1.73	9–16	0.65
G4	30	11.23	3.87	4–24	0.72
G5	30	14.53	7.02	6–44	0.63

As shown, sentence length and lexical diversity vary across categories. While grammatical baseline items (G2) and diastatically/diaphasically marked items (G5) tend to be longer on average, gray-zone items (G1) are comparatively more compact. Lexical diversity is highest in the diatopically marked condition (G4), consistent with the inclusion of regionally differentiated forms. No category exhibits extreme structural divergence that would suggest trivial length- or repetition-based confounds.

2.5. Data Collection

All prompts were submitted manually between 15 May and 20 May 2025, with each model receiving the full set of 150 items. Each response was logged with the following metadata:

- Model name and version;
- Prompted input;
- Model output (raw and parsed);
- Timestamp.

An overview of the data-collection and preprocessing pipeline is provided in Figure 1 below. The responses were compiled into a structured spreadsheet and cross-checked for format consistency. No model was exposed to the same item more than once. The present design therefore does not include repeated submissions of identical stimuli to assess intra-model rating stability across trials. Deployed LLM systems are generally non-deterministic: even under identical prompts, outputs may vary across runs due to stochastic decoding procedures, provider-side inference settings that are not fully exposed in web interfaces (e.g., temperature or routing mechanisms), and unannounced updates to deployed model versions. Because the present dataset contains only one observation per model $\times$ item, it does not allow estimation of within-model run-to-run variance, test–retest reliability, or uncertainty intervals around mean ratings. The analyses reported here therefore describe the observed behavior of the publicly accessible systems during the specified data-collection window rather than stable, fully reproducible properties of a fixed model checkpoint. Evaluating rating consistency across repeated runs—particularly for

marginal (G1) and variation-sensitive (G4/G5) items—would allow direct estimation of stochastic run-to-run variability and is therefore an important direction for future research.

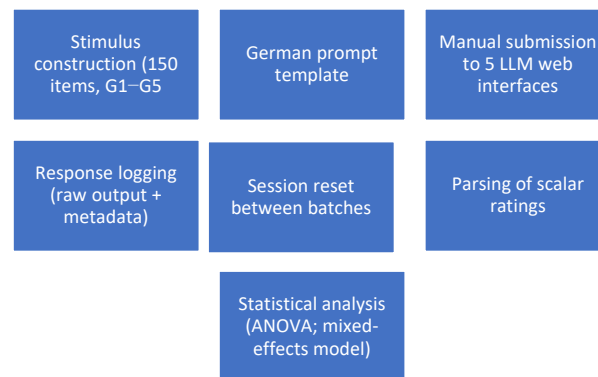


Figure 1. Workflow of data collection and preprocessing. The figure schematically illustrates the sequence of stimulus construction, prompting, manual submission, session management, logging, parsing, and statistical analysis.

To ensure data integrity, all inputs were submitted in isolated sessions, with conversation histories cleared between batches. Model outputs were documented immediately upon generation and preserved in their original form before any preprocessing or analysis. Parsing scripts were used to extract the scalar rating from each model output, along with any associated explanation, comment, or justification provided by the model. These justifications were stored for supplementary qualitative analysis.

The outputs were reviewed for anomalies such as rating omissions, invalid scale values or output truncation. In fewer than 2% of cases, incomplete or malformed responses were observed (e.g., where a rating was missing or ambiguous). These were flagged and re-submitted using identical prompts under a fresh session, with the re-run version retained for analysis. No model was prompted to revise or clarify its responses post hoc.

In order to control for prompt-based variability, the entire 150-item prompt was submitted in identical format to each model. Where platform limitations made it necessary to divide the prompt into smaller chunks (e.g., character limits), careful segmentation was applied to preserve the original item order and formatting. These variations are documented in the Supplementary Materials.

An important methodological consideration concerns the question of reproducibility when working with deployed LLMs. Reproducibility in such settings is necessarily constrained by the dynamic nature of these systems. Providers may update model architectures, inference pipelines or sampling procedures without public versioning at the level of specific checkpoints, and LLM outputs may exhibit stochastic variability even under identical prompts. The present study therefore documents model behavior as observed within a clearly specified data-collection window and under explicitly described access conditions. Exact replication of every individual rating cannot be guaranteed in future re-runs given the non-deterministic and evolving character of current LLM deployments. However, procedural reproducibility is ensured through the detailed documentation of all stimuli, prompt templates, logging procedures, and analytical steps in the manuscript. In this sense, the study provides a transparent and time-bound empirical record of system behavior rather than a claim of immutable model responses.

3. Results

This section presents the quantitative results of the pilot study evaluating LLMs as acceptability raters in German. As noted above, the analyses are based on scalar acceptability judgments elicited on a five-point Likert scale, where higher values correspond

to higher perceived acceptability. Each of the 150 stimulus sentences received exactly one rating from each of the five models, yielding a balanced dataset of 750 observations (5 models $\times$ 150 items). For the complete LLM-by-LLM ratings for all 150 stimuli, see the Data Availability Statement.

Although Likert-scale ratings are strictly ordinal, they were numerically encoded (1–5) and treated as approximately interval-scaled for the purposes of the omnibus analyses reported here. This analytical choice reflects the exploratory character of the study and follows established practice in experimental linguistics, where Likert-based acceptability judgments are commonly analyzed using parametric methods when scale points are equidistant and sample sizes are sufficiently large. In the present balanced design (5 models $\times$ 150 items), the primary focus lies on relative differences across models and stimulus categories rather than on a fine-grained metric interpretation of scale distances. The main contrasts reported below (e.g., grammatical vs. ungrammatical items) are robust to reasonable assumptions about scale structure. At the same time, it is acknowledged that the present design involves repeated measurements over shared stimulus items, and that more fine-grained analyses may benefit from statistical models that explicitly account for item-level dependencies and the ordinal nature of the response variable.

All statistical tests were evaluated at a significance level of $\alpha = 0.05$. The analyses were performed in JASP ([33]; <https://jasp-stats.org/>, accessed on 27 January 2026), an open-source software environment for statistical analysis.

3.1. Overall Model Effects on Acceptability Ratings

To address RQ1, we first report a descriptive baseline comparison of the models' overall mean rating levels, collapsing across all stimulus types. This analysis is intended to assess whether the models differ in general rating strictness or leniency when all 150 items are aggregated. It does not aim to model the cross-classified structure of the dataset (each item rated once by each model) or systematic differences across stimulus categories (G1–G5). Structure-sensitive inferential analyses that account for item-level variance are reported in Section 3.3.

For descriptive purposes, we performed a one-way ANOVA with model (ChatGPT-4, Gemini 2.0 Flash, Claude 3.5 Sonnet, Perplexity AI, DeepSeek) as a factor and acceptability rating as the dependent variable. In this analysis, ratings were collapsed across all stimulus categories, including canonical grammatical items, clearly ungrammatical constructions, gray-zone items with documented variation, as well as diatopically and diastratically/diaphasically marked utterances. The purpose of this aggregation was to test for global differences in mean rating levels between models, independent of specific linguistic phenomena.

The ANOVA revealed no statistically significant effect of model on acceptability ratings, $F(4, 745) = 0.846$, $p = 0.497$ (Table 4). The proportion of explained variance was negligible, indicating that differences between models account for virtually none of the total variance in aggregated acceptability ratings.

Table 4. Results of the one-way ANOVA testing for overall differences in acceptability ratings across LLMs.

ANOVA evaluation	Sum of Squares	df	Mean Square	F	p
AI model	8.061	4	2.015	0.846	0.497
Residuals	1775.693	745	2.383		

The absence of a significant omnibus effect indicates that, when ratings are averaged across all 150 stimuli, the five models exhibit comparable global acceptability baselines. In

other words, the models do not differ systematically in how leniently or strictly they assign acceptability ratings overall.

Importantly, this result should not be interpreted as evidence that the models behave identically across linguistically relevant dimensions. Rather, it reflects the fact that differences between models are not detectable at the level of grand means, where ratings for highly acceptable items and clearly unacceptable items are pooled together.

Given the composition of the stimulus set, this outcome is consistent with a scenario in which all models:

- assign uniformly high ratings to uncontroversially grammatical sentences;
- assign uniformly low ratings to clearly ungrammatical constructions;
- but diverge in their treatment of items involving morphosyntactic variation, register sensitivity, or sociolinguistic marking.

When such divergent behaviors occur in opposite directions across different subsets of stimuli, they may cancel each other out in aggregated analyses, resulting in statistically indistinguishable overall means. The present ANOVA therefore establishes an important baseline result: any substantive model-specific differences must emerge at a more fine-grained level of analysis, such as comparisons across stimulus categories, construction types, or sociolinguistic conditions.

Taken together, the results indicate that the overall distribution of acceptability ratings does not differ significantly across models when responses are aggregated across all stimulus types. This finding suggests that the five LLMs operate with broadly similar global notions of acceptability when evaluated in the abstract. At the same time, the null effect motivates subsequent analyses that preserve the internal structure of the stimulus set and are better suited to capturing gradient and context-sensitive aspects of linguistic acceptability. It is of course important to note that the above ANOVA treats ratings as independent observations and collapses across heterogeneous stimulus categories. While this is appropriate for assessing global mean-level differences, it does not explicitly model item-level variance or the fully crossed repeated-measures structure of the dataset. To address this, mixed-effects analyses that account for item-level variability and stimulus category are reported below.

3.2. Pairwise Comparisons Between Individual Language Models

To address RQ2, which asks whether statistically significant differences in acceptability judgments emerge between individual LLMs, post hoc pairwise comparisons were conducted following the omnibus ANOVA reported in Section 3.1. While the ANOVA did not reveal a significant overall effect of model on acceptability ratings, post hoc tests were nonetheless examined in order to assess whether any specific model pairs exhibit localized differences that might not be detectable at the level of the global test.

Pairwise comparisons between all five models were carried out using Tukey's Honestly Significant Difference (HSD) procedure, which controls the family-wise error rate when multiple comparisons are performed. For completeness, *p*-values adjusted using alternative correction methods (Bonferroni–Holm method; [34]) were also inspected. The comparisons tested differences in mean acceptability ratings aggregated across all 150 stimuli, mirroring the aggregation strategy employed in the omnibus analysis.

The results of the post hoc tests are summarized in Table 5. Across all model pairs, none of the mean differences reached statistical significance, regardless of the correction method applied. This holds even before correction for multiple comparisons: unadjusted test statistics already indicate that none of the observed differences approach conventional significance thresholds:

Table 5. Post hoc pairwise comparisons of LLMs based on mean acceptability ratings (Tukey HSD).

Post Hoc Tests								
Standard (HSD)								
Post Hoc Comparisons–AI model								
		Mean difference	SE	df	t	P _{Tukey}	P _{bonf}	P _{holm}
(ChatGPT-4)	Claude 3.5 Sonnet	0.067	0.178	745	0.374	0.996	1.000	1.000
	Deep Seek	−0.053	0.178	745	−0.299	0.998	1.000	1.000
	Gemini 2.0 Flash	−0.233	0.178	745	−1.309	0.686	1.000	1.000
	Perplexity AI	−0.127	0.178	745	−0.711	0.954	1.000	1.000
Claude 3.5 Sonnet	Deep Seek	−0.120	0.178	745	−0.673	0.962	1.000	1.000
	Gemini 2.0 Flash	−0.300	0.178	745	−1.683	0.445	0.928	0.928
	Perplexity AI	−0.193	0.178	745	−1.085	0.815	1.000	1.000
Deep Seek	Gemini 2.0 Flash	−0.180	0.178	745	−1.010	0.851	1.000	1.000
	Perplexity AI	−0.073	0.178	745	−0.411	0.994	1.000	1.000
Gemini 2.0 Flash	Perplexity AI	0.107	0.178	745	0.598	0.975	1.000	1.000

Note: *p*-value adjusted for comparing a family of 5 estimates.

The estimated mean differences between models are uniformly small in magnitude, ranging approximately between −0.30 and +0.11 on the five-point Likert scale. These differences correspond to less than one third of a scale point and thus reflect minimal practical divergence in overall rating behavior. Standard errors are comparable across contrasts, and the associated test statistics ($|t| < 1.7$ in all cases) further indicate the absence of systematic separation between model means.

Notably, this pattern holds across all theoretically relevant pairings, including comparisons between models developed by different providers (e.g., ChatGPT-4 vs. Gemini 2.0 Flash, ChatGPT-4 vs. Claude 3.5 Sonnet) as well as between more established and more recently introduced systems (e.g., Claude 3.5 Sonnet vs. DeepSeek). Even contrasts that might be expected to differ due to architectural, training, or deployment differences do not exhibit reliable mean-level differences in acceptability ratings.

Taken together, the post hoc results corroborate the findings of the omnibus ANOVA by demonstrating that no pair of models differs significantly from any other in terms of their overall mean acceptability judgments. This convergence across multiple correction methods strengthens the conclusion that the absence of statistically significant effects is not an artifact of conservative adjustment procedures but rather reflects genuinely small between-model differences at the level of aggregated ratings.

Importantly, this result does not imply that the models behave identically in all respects. Rather, it suggests that global mean ratings are not a sensitive measure for capturing differences in how LLMs respond to linguistically complex or variation-sensitive stimuli. As with the omnibus analysis, aggregating across heterogeneous stimulus types is likely to obscure localized divergences in model behavior that arise in specific linguistic domains.

Thus, the post hoc comparisons reinforce a key methodological insight of the present study: model-specific differences, where they exist, are unlikely to manifest as simple shifts in overall acceptability level but instead require analyses that preserve the internal structure of the stimulus set, including distinctions between grammaticality, variation, register, and sociolinguistic marking.

In summary, in response to RQ2 the results indicate that there are no statistically significant differences in aggregated acceptability judgments between any pair of evaluated

LLMs. All models exhibit highly similar overall rating profiles when responses are collapsed across stimuli, and none of the observed mean differences survives correction for multiple comparisons. This finding complements the results reported for RQ1 and further motivates subsequent analyses that move beyond global averages to examine model behavior in linguistically structured subsets of the data.

3.3. Mixed-Effects Analysis Accounting for Item-Level Variance

While the grand-mean ANOVA reported in Section 3.1 provides a descriptive baseline comparison of overall rating levels, the dataset exhibits a fully crossed repeated-measures structure: each of the 150 items was rated once by each of the five models. In such a design, ratings are not independent observations, as variability attributable to individual stimulus items may systematically influence the results. In addition, the stimulus set is internally structured by category (G1–G5), introducing theoretically motivated heterogeneity.

To account for the fully crossed repeated-measures structure of the dataset, a linear mixed-effects model was fitted with acceptability rating as the dependent variable. As noted by an anonymous reviewer, potential differences between models may only become visible once variation across stimulus categories is taken into account. To address this point, the mixed-effects model was extended to include an interaction between Model and Category. Model, Category and their interaction were entered as fixed effects. Item was included as a random intercept to capture item-specific baseline variability and to account for the non-independence of observations arising from each item being rated by all models.

To examine whether model differences vary across linguistic domains, the mixed-effects model included an interaction between Model and Category. Several interaction terms reached statistical significance. In particular, strong interaction effects emerged for the ungrammatical condition (G3). Relative to ChatGPT-4, all other models assigned significantly higher ratings to ungrammatical sentences (Claude: $\beta = 0.833$, $p < 0.001$; DeepSeek: $\beta = 1.067$, $p < 0.001$; Gemini: $\beta = 0.667$, $p < 0.001$; Perplexity: $\beta = 0.967$, $p < 0.001$). This indicates that ChatGPT-4 penalizes morphosyntactic violations more strongly than the other evaluated systems. Smaller interaction effects were observed in other categories. For grammatical baseline items (G2), Claude 3.5 Sonnet produced slightly higher ratings than ChatGPT-4 ($\beta = 0.367$, $p = 0.029$). For diatopically marked items (G4), Claude again showed a modest increase in ratings relative to ChatGPT-4 ($\beta = 0.467$, $p = 0.005$). In the diastratically/diaphasically marked condition (G5), DeepSeek assigned somewhat higher ratings than ChatGPT-4 ($\beta = 0.333$, $p = 0.047$). These results indicate that differences between models do not reflect a uniform global shift in rating behavior but vary depending on the linguistic domain under consideration.

The results of the mixed-effects model are summarized in Table 6. Test statistics are reported as Wald z-values as provided by JASP.

The variance attributable to Item was 0.93, indicating substantial item-level heterogeneity in baseline acceptability ratings. This confirms that a non-negligible portion of variance is associated with item-specific baseline differences, justifying the use of a mixed-effects approach.

As expected, a strong main effect of Category emerged, reflecting the intended differentiation between grammatical, ungrammatical, and variation-sensitive stimulus types. In particular, ungrammatical items (G3) were rated significantly lower than gray-zone items (reference level), whereas grammatical baseline items (G2) and diatopically marked items (G4) received significantly higher ratings (all $p < 0.001$). This confirms that the categorical manipulation of stimulus type robustly structures the rating distribution.

Table 6. Results of the linear mixed-effects model predicting acceptability ratings (random intercept for Item). Fixed effects include Model, Category and their interaction.

Effect	β	SE	z	p
Intercept (ChatGPT-4, G1)	3.867	0.195	19.83	<0.001
Model: Claude 3.5 Sonnet	−0.433	0.119	−3.65	<0.001
Model: DeepSeek	−0.233	0.119	−1.97	0.049
Model: Gemini 2.0 Flash	0.167	0.119	1.40	0.160
Model: Perplexity AI	−0.133	0.119	−1.12	0.262
Category: G2	1.133	0.276	4.11	<0.001
Category: G3	−2.500	0.276	−9.06	<0.001
Category: G4	1.100	0.276	3.99	<0.001
Category: G5	−0.533	0.276	−1.93	0.053
Model $\times$ Category (overall interaction)	-	-	$\chi^2 = 54.3$	<0.001

With respect to Model, there was no evidence of a large uniform shift in acceptability ratings across categories that would indicate a consistent overall leniency or strictness difference between systems. Although some individual model coefficients relative to the reference level (ChatGPT-4) reached statistical significance, their magnitude was small (all $|\beta| < 0.5$ on a five-point scale), indicating only minor deviations in baseline rating behavior.

Collectively, the mixed-effects analysis complements the omnibus ANOVA: while the models share broadly similar grand-mean baselines, category structure accounts for substantial variance and the interaction analysis indicates greater differentiation in variation-sensitive conditions. The absence of a significant grand-mean model effect should therefore not be interpreted as uniformity across all stimulus domains but rather as evidence that differences are category-specific rather than global.

At first glance, the mixed-effects results may appear to differ from the omnibus ANOVA reported in Section 3.1, which did not detect a significant overall effect of Model. This difference arises from the statistical structure of the two analyses. The ANOVA treats all observations as independent and collapses across item-level variability, whereas the mixed-effects model explicitly accounts for variance attributable to individual stimulus items. By modeling item-specific intercepts, the mixed-effects framework removes a portion of unexplained variance and therefore increases sensitivity to small systematic differences between models. As a result, minor baseline deviations between systems become detectable in the mixed-effects analysis even though they do not produce a significant global effect in the aggregated ANOVA.

3.4. Category-Specific Analysis

To obtain a more fine-grained understanding of model behavior, mean acceptability ratings were calculated separately for each stimulus category (G1–G5). Table 7 summarizes the category-specific means for all five models:

Across grammatical baseline items (G2), all models show near-ceiling performance (means between 4.83 and 5.00), indicating consistent recognition of well-formed standard constructions. Conversely, ungrammatical items (G3) received uniformly low ratings (means between 1.37 and 2.20), demonstrating reliable detection of clear syntactic violations across systems. Gray-zone items (G1) yielded intermediate ratings (means between 3.43 and 4.03), reflecting their intended gradient status. Here, model differences

are somewhat more pronounced, suggesting varying sensitivity to marginal or morphosyntactically innovative constructions. For diatopically marked items (G4), all models assigned very high ratings (means between 4.57 and 5.00). Regionally marked forms were thus largely treated as acceptable rather than erroneous. This indicates that dialectal variation is generally distinguished from grammatical violation by the models. In contrast, diastratically and diaphasically marked items (G5) show more moderate ratings (means between 2.97 and 3.43). Compared to G4, stylistic and register-based deviations appear to be evaluated more cautiously. While colloquial or informal constructions are often accepted, strongly marked register shifts and stylistic exaggerations tend to lower ratings. This suggests that register variation elicits greater evaluative variability than diatopic variation.

Table 7. Mean acceptability ratings by model and stimulus category (tabular).

Model	G1	G2	G3	G4	G5
ChatGPT-4	3.87	5.00	1.37	4.97	3.33
Gemini 2.0 Flash	4.03	4.93	2.20	5.00	3.40
Claude 3.5 Sonnet	3.43	4.93	1.77	5.00	2.97
Perplexity AI	3.73	4.83	2.20	5.00	3.33
DeepSeek	3.63	4.97	2.20	4.57	3.43

As a whole, the category-specific analysis refines the global results by indicating that model behavior is highly consistent for clear grammaticality contrasts (G2 vs. G3), while variation-sensitive categories (G1 and G5) reveal greater differentiation. The category-specific mean ratings are visualized in Figure 2:

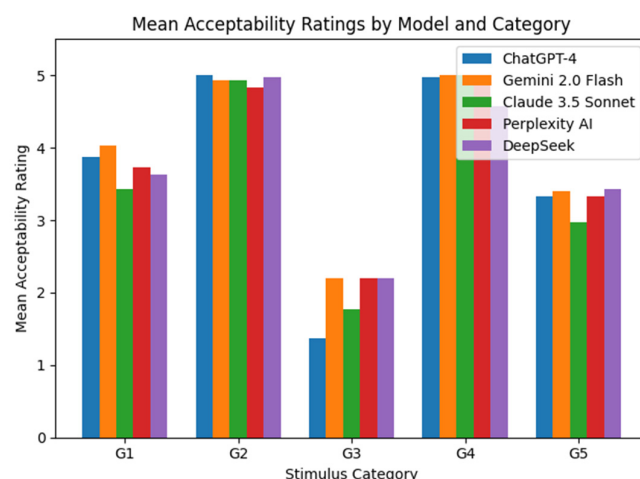


Figure 2. Mean acceptability ratings by model and stimulus category (bar chart).

Figure 2 displays the category-specific mean ratings summarized in Table 7 for all five evaluated LLMs. A clear separation between grammatical (G2) and ungrammatical (G3) items is observed across models, while gray-zone (G1) and diastratically/diaphasically marked items (G5) exhibit greater variability in ratings.

4. Discussion

The analyses presented in Section 3 provide a baseline for interpreting LLM behavior in gradient acceptability judgment tasks in German. Rather than revealing differences between models at the level of aggregated mean ratings, the results raise questions about how and where potential differences between models are likely to emerge once linguistically

relevant structure is taken into account. In the following, these questions are discussed with reference to the variation-sensitive design of the study and to current debates on the role of LLMs in linguistic research.

4.1. Interpreting the Absence of Global Model Effects

The non-significant omnibus ANOVA and uniformly non-significant post hoc contrasts indicate that when acceptability ratings are collapsed across all stimulus types, the five evaluated models do not differ systematically in their overall rating behavior. In this sense, differences in provider, deployment conditions, or model family do not translate into simple shifts in scalar judgment behavior at the level of grand means.

At the same time, this result must be interpreted with caution. The stimulus set deliberately combines highly acceptable standard constructions, categorically ungrammatical sentences, and items whose acceptability depends on sociolinguistic, regional, or stylistic conditions. Aggregating across such heterogeneous domains necessarily reduces the sensitivity of global statistics to localized divergences. It is therefore methodologically expected that potential differences between models would not surface as uniform mean shifts but rather as patterned variation tied to specific linguistic domains. The mixed-effects analysis reported in Section 3.3 confirms this expectation: model differences do not appear as global rating shifts, but instead emerge in specific stimulus categories, particularly in the evaluation of ungrammatical constructions.

The present null effect thus does not license the conclusion that the models behave identically in all respects. Instead, it highlights the limited interpretive value of global mean comparisons in acceptability research and underscores the importance of analyses that preserve distinctions between grammaticality, variation, and contextual appropriateness.

4.2. Acceptability, Normativity and Variation as Analytical Dimensions

The design of the study reflects a view of acceptability as a multi-dimensional phenomenon rather than as a purely grammatical property. Accordingly, the stimulus set includes not only categorical morphosyntactic contrasts but also constructions whose evaluation depends on register, context, and sociolinguistic marking.

From this perspective, the inclusion of gray-zone items, diatopically marked constructions and diastratically/diaphasically marked expressions serves not merely to increase empirical coverage, but to test whether LLMs respond to the same multi-layered constraints that shape human judgments. The study design therefore targets precisely those domains in which surface well-formedness and contextual appropriateness come apart, such as register mismatches or regionally restricted constructions that are grammatical but socially or stylistically marked.

The fact that such phenomena are not separable from broader issues of normativity is particularly relevant. Linguistic norms are neither uniform nor static but distributed across competing standards, local practices, and socially indexed forms. Whether and how LLMs can represent this multiplicity of norms is thus a central question for their use as linguistic evaluators.

4.3. Why Variation-Sensitive Acceptability Is Challenging for LLMs

The variation-sensitive dimensions targeted in this study raise specific challenges for current LLM architectures and training regimes. Categorical morphosyntactic violations (e.g., agreement or case mismatches or unfelicitous word orders) tend to be consistently penalized across contexts and speaker groups and are therefore strongly represented as low-probability patterns in large text corpora. In contrast, constructions involving variation, register sensitivity or regional marking presuppose the coexistence of competing

norms whose acceptability is conditioned by discourse context, speaker identity and audience design.

LLMs trained on heterogeneous and predominantly edited corpora are structurally biased toward globally frequent and standardized patterns. As a consequence, such models may privilege prescriptively dominant or unmarked variants when faced with norm competition, even in cases where non-standard or contextually marked alternatives are locally appropriate. While traces of variation are undoubtedly present in training data, they are not explicitly modeled as socially anchored alternatives within a structured system of linguistic norms.

This provides a plausible explanation for tendencies—already suggested by the design of the present study and by informal or anecdotal inspection of selected model outputs—to normalize surface variation, default to standard-like forms, or respond categorically to inputs intended to elicit graded judgments. Importantly, such behavior need not reflect random error but may instead arise from systematic biases toward dominant norm regimes and from limited access to the sociolinguistic metadata that human speakers routinely exploit in acceptability judgments. These observations are exploratory and were not based on a systematic qualitative coding of model explanations.

4.4. Methodological Implications for LLM-Based Acceptability Research

The present findings bear directly on the methodological use of LLMs in acceptability research. On the one hand, the convergence of models in their treatment of uncontroversially grammatical and clearly ungrammatical constructions suggests that LLMs can function as useful exploratory tools in domains governed by relatively stable morphosyntactic constraints. In such areas, model judgments may serve as efficient sanity checks, aids in hypothesis generation, or preliminary probes prior to human experimentation.

On the other hand, the variation-aware design of the present study highlights important limitations. In domains shaped by norm competition, register sensitivity, and sociolinguistic conditioning, LLM judgments risk systematically overrepresenting dominant, standardized, or prescriptively privileged patterns at the expense of locally valid or socially meaningful alternatives. Accordingly, LLMs should not be treated as substitutes for speaker data in variationist or sociolinguistically oriented research but rather as heuristic instruments whose biases themselves can become objects of linguistic analysis.

From this perspective, the value of LLMs in linguistics does not lie solely in their potential to approximate speaker intuitions, but also in their capacity to reveal which dimensions of linguistic knowledge are readily captured by large-scale distributional learning—and which remain resistant to such modeling.

4.5. Limitations and Directions for Future Research

As a pilot study, the present investigation is subject to limitations that also point toward avenues for future research. First, the analyses reported in Section 3 focus on aggregated ratings as a baseline test of global model differences. However, the study's core hypothesis—namely, that differences may be more apparent in variation-sensitive domains—calls for follow-up analyses comparing models across stimulus categories and construction types, as well as examining interactions between model, linguistic phenomenon, and contextual manipulation.

Second, Likert-scale ratings were treated as approximately interval-scaled for the purposes of omnibus testing here. However, future analyses should ideally employ ordinal mixed-effects models (e.g., cumulative link mixed models) that account for the ordinal nature of the rating scale and the ordinal nature of the response variable, such as cumulative link mixed models (CLMMs) or other ordinal mixed-effects regression approaches.

In addition, all systems were queried via publicly accessible web interfaces using provider-defined default inference settings. Because decoding parameters such as temperature and sampling strategies are not standardized across platforms and are not user-configurable in all cases, cross-model differences may partly reflect deployment-level inference settings rather than underlying architectural properties. This constitutes an additional source of uncontrolled variance in the present design.

Finally, one of the most promising—though methodologically challenging—directions for future research is a comparison with human judgments on the same dataset. As noted above, this is not addressed here for reasons of space but forms part of a larger project by one of the authors (see also [15] for results from a further pilot study).

Considered together, these considerations position the present study as a methodological and empirical foundation for more fine-grained investigations into LLM behavior in linguistically structured domains.

5. Conclusions

This paper presented a pilot study comparing five LLMs as raters in a gradient acceptability judgment task in German, using a contextually embedded stimulus set spanning grammatical, ungrammatical, variation-sensitive, and sociolinguistically marked constructions.

The baseline quantitative analyses revealed no statistically significant overall model effect on aggregated acceptability ratings. Both the omnibus ANOVA and post hoc comparisons indicate that, when responses are collapsed across stimulus types, the evaluated systems operate with comparable global acceptability baselines.

However, convergence at the level of grand means does not entail uniform behavior across linguistically structured domains. Acceptability is shaped not only by morphosyntax but also by normativity, register, and sociolinguistic conditioning. Analyses based solely on aggregated ratings are therefore limited in their ability to capture domain-specific differences in model behavior. The variation-aware design adopted here provides a framework for examining precisely those dimensions in which grammatical structure and social norms intersect.

Beyond its methodological implications, the findings also bear on the conceptual question of how linguistic normativity is represented in large language models. In variation-sensitive domains, model judgments appear to reflect not only structural well-formedness but also the relative dominance of standardized usage patterns in training data. This suggests a tendency toward norm consolidation, whereby socially or regionally indexed alternatives may be evaluated more cautiously than canonical forms. In this respect, LLMs function not only as heuristic acceptability raters but also as empirical indicators of how normative hierarchies and sociolinguistic diversity are encoded within large-scale distributional representations of language.

From a methodological perspective, the results support a differentiated use of LLMs in acceptability research. The systems evaluated here appear robust with respect to core grammatical contrasts, while their handling of variation-sensitive and context-dependent phenomena remains an open empirical question. Future work will extend the present design through category-specific analyses, statistical models that more explicitly reflect the structure of acceptability data, and systematic comparison with human judgments. In this way, the study contributes an empirically transparent baseline for assessing the role of LLMs in experimental and theoretical linguistics.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/ai7030112/s1>, Document S1: Stimulus set and ratings.

Author Contributions: Conceptualization, methodology, validation, formal analysis, investigation, resources, writing and original draft preparation, review and editing, funding acquisition: N.C.; data curation: N.C. and F.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the BUW Individuell (FORIS) program (Bergische Universität Wuppertal) during the period 1 November 2024–31 October 2025.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the Supplementary Materials. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Chomsky, N. *Aspects of the Theory of Syntax*; MIT Press: Cambridge, MA, USA, 1965.
- Chomsky, N. *Knowledge of Language: Its Nature, Origin and Use*; Praeger: New York, NY, USA, 1986.
- Levelt, W.J.M.; van Gent, J.A.W.M.; Haans, A.F.J.; Meijers, A.J.A. Grammaticality, paraphrase, and imagery. In *Acceptability in Language*; Greenbaum, S., Ed.; Mouton: The Hague, The Netherlands, 1977; pp. 87–101. [\[CrossRef\]](#)
- Featherston, S. Experimentell erhobene Grammatikalitätsurteile und ihre Bedeutung für die Syntaxtheorie. In *Sprachkorpora. Datenmengen und Erkenntnisfortschritt*; Kallmeyer, W., Zifonun, G., Eds.; de Gruyter: Berlin, Germany, 2007; pp. 49–69. [\[CrossRef\]](#)
- Schütze, C.T. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*; Language Science Press: Berlin, Germany, 2016. [\[CrossRef\]](#)
- Sprouse, J. Acceptability judgments and grammaticality, prospects and challenges. In *Syntactic Structures After 60 Years*; Hornstein, N., Lasnik, H., Patel-Grosz, P., Yang, C., Eds.; de Gruyter: Berlin, Germany, 2018; pp. 195–223. [\[CrossRef\]](#)
- Grice, H.P. Logic and conversation. In *Syntax and Semantics*; Cole, P., Morgan, J., Eds.; Academic Press: New York, NY, USA, 1975; pp. 41–58. [\[CrossRef\]](#)
- van Dijk, T.A. *Text and Context: Explorations in the Semantics and Pragmatics of Discourse*; Longman: London, UK, 1977.
- Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K. Deep contextualized word representations. In *Proceedings of NAACL-HLT 2018*; ACL: New Orleans, LA, USA, 2018; pp. 2227–2237.
- Qiu, Z.; Duan, X.; Cai, Z. Evaluating grammatical well-formedness in large language models: A comparative study with human judgments. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*; ACL: Bangkok, Thailand, 2024; pp. 189–198.
- Hu, J.; Mahowald, K.; Lupyan, G.; Ivanova, A.; Levy, R. Language models align with human judgments on key grammatical constructions. *Proc. Natl. Acad. Sci. USA* **2024**, *121*, e2400917121. [\[CrossRef\]](#) [\[PubMed\]](#)
- Leivada, E.; Günther, F.; Dentella, V. Reply to Hu et al.: Applying different evaluation standards to humans vs. Large Language Models overestimates AI performance. *Proc. Natl. Acad. Sci. USA* **2024**, *121*, e2406752121. [\[CrossRef\]](#) [\[PubMed\]](#)
- Johnsen, L.G.B. Grammaticality judgments in humans and language models: Revisiting Generative Grammar with LLMs. *arXiv* **2025**, arXiv:2512.10453. [\[CrossRef\]](#)
- Ide, Y.; Nishida, Y.; Vasselli, J.; Oba, M.; Sakai, Y.; Kamigaito, H.; Watanabe, T. How to make the most of LLMs' grammatical knowledge for acceptability judgments. In *Proceedings of NAACL-HLT 2025*; ACL: Miami, FL, USA, 2025; pp. 7416–7432.
- Catasso, N. Benchmarking AI acceptability and grammaticality in German: A study of ChatGPT and human judgments. *AI Linguist.* **2026**, *3*, 1–41. [\[CrossRef\]](#)
- OpenAI. ChatGPT (GPT-4) [Large Language Model]. Available online: <https://chat.openai.com> (accessed on 12 January 2026).
- Google DeepMind. Gemini 2.0 Flash [Large Language Model]. Available online: <https://gemini.google.com> (accessed on 12 January 2026).
- Anthropic. Claude 3.5 Sonnet [Large Language Model]. Available online: <https://claude.ai> (accessed on 12 January 2026).
- Perplexity AI. Perplexity AI Platform [LLM-Based System]. Available online: <https://www.perplexity.ai> (accessed on 12 January 2026).
- DeepSeek. DeepSeek Chat [Large Language Model]. Available online: <https://chat.deepseek.com> (accessed on 12 January 2026).
- Klein, W.P. (Ed.) *Sprachliche Zweifelsfälle: Theorie und Empirie/Grammatical Uncertainties: Theoretical and Empirical Aspects*; Linguistik Online; Bern Open Publishing: Bern, Switzerland, 2003. [\[CrossRef\]](#)
- Klein, W.P. *Sprachliche Zweifelsfälle im Deutschen. Theorie, Praxis, Geschichte*; de Gruyter: Berlin, Germany; Boston, MA, USA, 2018.
- Vikner, S. Immobile complex verbs in Germanic. *J. Comp. Ger. Linguist.* **2005**, *8*, 83–115. [\[CrossRef\]](#)

24. Freywald, U.; Simon, H.J. Wenn die Wortbildung die Syntax stört: Über Verben, die nicht in V2 stehen können. In *Verbale Wortbildung im Spannungsfeld zwischen Wortsemantik, Syntax und Rechtschreibung*; Kauffer, M., Métrich, R., Eds.; Stauffenburg: Tübingen, Germany, 2007; pp. 181–194.
25. Forche, C.R. *NonV2-Verben im Deutschen: Theoretische Überlegungen und Empirische Untersuchungen zu Einem Morphosyntaktischen Problemfall (Den es Vielleicht gar Nicht Gibt)*; Springer: Berlin/Heidelberg, Germany, 2020. [CrossRef]
26. DWDS—Digitales Wörterbuch der Deutschen Sprache. Entry: *Bauchreden*. Available online: <https://www.dwds.de/wb/bauchreden> (accessed on 12 December 2025).
27. Duden Online. Entry: *Bauchreden*. Available online: <https://www.duden.de/rechtschreibung/bauchreden> (accessed on 1 December 2025).
28. Fleischer, J. *Die Syntax von Pronominaladverbien in den Dialekten des Deutschen*; Steiner: Stuttgart, Germany, 2002.
29. Fleischer, J. Preposition stranding in German dialects. In *Syntactic Microvariation*; Barbiers, S., Cornips, L., van der Kleij, S., Eds.; Mertens Instituut: Amsterdam, The Netherlands, 2002; pp. 116–151.
30. Leser, S. Zum Pronominaladverb in den hessischen Dialekten. In *Niederdeutsche Syntax*; Langhanke, R., Berg, K., Elmentaler, M., Peters, J., Eds.; Olms: Hildesheim, Germany, 2012; pp. 79–99.
31. Negele, M. *Varianten der Pronominaladverbien im Neuhochdeutschen*; de Gruyter: Berlin, Germany, 2012. [CrossRef]
32. Cirkel, P.; Freywald, U. In Stadt und Stadt: Berlin und Ruhrgebiet im Vergleich. *Linguist. Online* **2021**, *110*, 193–227. [CrossRef]
33. JASP Team. JASP (Version 0.95.1) [Computer Software]. Available online: <https://jasp-stats.org> (accessed on 1 January 2026).
34. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.