Assessing the Effect of Context in Multi-domain Acceptability Judgment

Eunike Andriani Kardinata, Yusuke Sakai, Taro Watanabe
Nara Institute of Science and Technology (NAIST), Japan
{eunike.kardinata.ef9, sakai.yusuke.sr9, taro}@is.naist.jp

Abstract

Acceptability judgments provide a crucial basis for understanding how sentences are perceived as natural or well-formed, and they are increasingly used to assess the linguistic capability of large language models (LLMs). Unlike grammaticality, acceptability depends not only on structural form but also on contextual and domain-specific factors. Most prior work evaluates sentences in isolation, and relatively little is known about how explicit contextual cues influence LLM acceptability judgments across domains. This study examines how contextual information affects model-generated acceptability ratings across multiple domains and several LLMs, using different forms of domain-specific contextual cues to situate sentences in their intended usage settings. The results show that context can meaningfully shift model judgments, although its effects vary across models and domains. Overall, the findings provide evidence on contextual effects in LLM acceptability judgment and support the development of more context-aware evaluation frameworks.

1 Introduction

Research in theoretical linguistics treats acceptability judgments as behavioural evidence about speakers' grammatical knowledge and the relationship between grammatical structure and observed intuitions (Sprouse, 2023). Prior work emphasizes that acceptability is not identical to grammaticality (Juzek, 2024; Sprouse and Schütze, 2019). Unlike grammaticality, which concerns structural well-formed-ness, acceptability is further shaped by discourse expectations, domain conventions, and stylistic norms (Myers, 2017; Fanselow, 2021). As large language models (LLMs) are increasingly utilized in text evaluation and interpretation (Li et al., 2025; Bavaresco et al., 2025; Rudnicka, 2025), it has become essential to understand how their acceptability judgments are influenced by contexts rather than by sentence-internal properties alone.

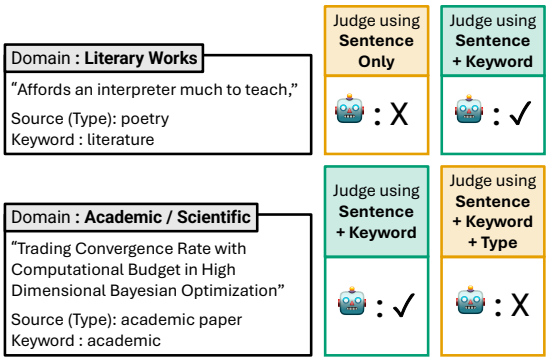

Figure 1: Variations in acceptability judgment

Existing studies evaluate sentences in isolation and gives limited attention to how context influences interpretation (Warstadt et al., 2019; Mikhailov et al., 2022; Someya et al., 2024; Beauchemin and Khoury, 2025). There is yet more to investigate about adjustment of LLM judgments given contextual cues in different forms and how it varies across different domains of language use. For example, consider the sentence in Figure 1, "Affords an interpreter much to teach." In the absence of contextual cues, this sentence is likely to be judged unacceptable, as it appears incomplete despite being grammatically well-formed. When contextual cues indicate that the sentence occurs in a literature–specifically, in poetry–the judgment may shift. In that setting, the sentence can be interpreted as stylistically appropriate rather than deviating from standard prose usage. This gap is thus especially important because many sentences are only fully interpretable when their functional setting or communicative purpose is taken into account (Bizzoni and Lappin, 2018; Sinha et al., 2023).

Our study addresses this gap by examining how explicit contextual cues influences LLM acceptability judgments across multiple domains, i.e., literature, journalistic, advertising, academic, instructional, communicative, online media, and legal. Not only conventional full-sentence inputs, we

Findings of the Association for Computational Linguistics: ACL 2026, pages 42251–42266

also consider short-form texts such as titles and advertisement snippets. We hypothesize that providing contextual cues facilitates a model's ability to determine whether a sentence is acceptable, particularly in short-form contexts where no surrounding discourse is available. Prior work has shown that selecting impactful keywords is effective for producing high-quality headlines (E et al., 2023), suggesting that keyword cues play a meaningful role in how models interpret textual content. Assuming that sentences are acceptable within their own domain, the proportion of acceptable judgments should increase relative to a baseline without context. More broadly, we contribute towards a context-aware perspective on model behaviour, one that aligns more closely with how language is encountered and interpreted in real-world settings.

Our findings show that contextual cues do not influence LLMs uniformly: its effects vary across domains and across models. In particular, models tend to respond more strongly to concrete, example-oriented cues than to abstract definitional information, and domains with atypical lexical, stylistic, or structural properties benefit most from contextual support. Moreover, models diverge in their acceptance thresholds and contextual sensitivity. Their sentence-level judgments become less aligned when context is introduced, indicating model-specific integrations of cues.

2 Related Work

A probabilistic perspective on the gradience of acceptability (Francis, 2021) is developed by Lau et al. (2017), and subsequent works show that contextual information and discourse environment can modulate both human and model acceptability ratings (Bernardy et al., 2018; Iavarone et al., 2021). Prior research has primarily examined complete sentences, introducing contextual information through prefixes (Sinha et al., 2023) or preceding sentences (Lau et al., 2020); however, these approaches are not well suited to short-form texts, where no surrounding discourse is available. Besides, existing studies typically characterize context only in terms of in-domain versus out-of-domain contrasts, rather than the more fine-grained functional domains. The distinction between grammaticality and acceptability is further supported by studies demonstrating systematic divergences between grammatical well-formed-ness and observed acceptability behaviour (Juzek, 2024).

A substantial body of methodological research examines how acceptability-judgment tasks should be designed and interpreted (Sprouse et al., 2013; Juzek, 2015; Sprouse and Almeida, 2017; Langsford et al., 2018). These research show that acceptability data are method-dependent and that careful experimental control is required when comparing conditions or interpreting rating differences. Within computational linguistics, acceptability has been operationalized as an evaluation task for neural language models (Warstadt et al., 2019; Vázquez Martínez, 2021) with additional analysis highlighting biases which substantially affects reported performance (Daultani et al., 2024).

Recent works also examine how different methods and application settings shape model-based acceptability judgment. Ide et al. (2025) compares multiple prompting and probability-based methods for deriving acceptability judgments from LLMs. Zhang et al. (2025) introduces the AdTEC benchmark for advertising text and evaluates model and human performance across multiple quality-related tasks, showing that performance varies across domains. These emphasize that model-based acceptability behaviour is sensitive to linguistic structure, evaluation method, genre, and application context.

3 Context in Multi-domain Acceptability Judgement

To examine how contextual cues shape model judgments, we define the basis of our proposed methodology. We assume that contextual cues help models assess sentence acceptability more effectively; given that a sentence is expected to be acceptable within its own domain, we anticipate a higher proportion of acceptable judgments compared to a baseline condition without context.

3.1 Task Definition

In this task, acceptability refers to how natural a sentence appears to a native speaker. While this notion may overlap with grammaticality, the two are not necessarily equivalent: a sentence may be judged acceptable even if it is not fully grammatical, and conversely, a grammatically well-formed sentence may be perceived as unacceptable in a given context. The rating is implemented as a binary decision, where **0** denotes an unacceptable sentence and **1** denotes an acceptable one.

The acceptability judgment is conducted under four distinct conditions, each corresponding to a

Domain : *Keyword*	Types	Definition
Literary Works (**lit**) : *literature*	poems, stories, novels, plays	(literature) written artistic works, especially those with a high and lasting artistic value
Journalistic (**jou**) : *journalistic*	news articles, features, reviews	(journalism): collections, writings, and publication of news stories and articles in newspapers and magazines
Advertisements (**adv**) : *advertisement*	slogans, ads, posters, promotional material	(advertisement): a piece of text that tries to persuade people to buy a product or service, or tells people about a job or event
Academic, Scientific (**acd**) : *academic*	research papers, reports, scientific essays	(research): detailed study of a subject, especially in order to discover new information about something or reach a new understanding of it
Instructional, Technical (**ins**) : *instructional*	manuals, guides, instructions	(manual): practical instructions on how to do something or how to use something, such as a machine
Functional, Everyday (**com**) : *communication*	emails, letters, forms, notices	(letter, notice): a written message from one person to another or a text containing information or instructions
Digital, Online (**onl**) : *online media*	blogs, social media posts, webpages	(blog): a record of thoughts, opinions, or experiences that is put on the internet for other people to read
Legal, Official Documents (**lgl**) : *legal*	contracts, laws, policies, case files	(contract, law): an official document that states and explains a formal agreement or a rule, usually made by a government, that is used to order the way in which a society behaves

Table 1: Domain specifications

different prompt. The conditions are: (1) Sentence Only, (2) Sentence + Domain Keyword (Keyword), (3) Sentence + Keyword + Types of Work (Types), (4) Sentence + Keyword + Domain Definition (Definition). In **Sentence Only (SO)**, the LLMs receive only the target sentence and are asked to judge its acceptability based solely on their internal linguistic knowledge and reasoning processes. This condition serves as a baseline, since no additional contextual cues might influence the judgment.

In the remaining three conditions, a domain keyword is introduced as an explicit contextual cue. In **Sentence + Keyword (SK)**, we assess whether the models adjust their ratings when minimal contextual cues are supplied. We then compare **Sentence + Keyword + Types (SKT)** and **Sentence + Keyword + Definition (SKD)** to examine which form of additional cue is more effective in guiding model behaviour—providing exemplar-like category instances (Types) or supplying a conceptual elaboration of the keyword (Definition). The contrast between SKT and SKD is intended to reveal whether models are more sensitive to either of the two when producing acceptability judgments.

3.2 Domains and Data

We defined eight distinct domains as described in Table 1 based on text function and intended usage. For each domain, the associated definition was derived from the Cambridge English Dictionary[1] for either the domain keyword or the relevant text types that provided the clearest explanatory value.

To construct the dataset for the experiment[2], we collected publicly available datasets from the web that were sufficiently representative of the eight domains under study. For each source dataset as specified in Appendix A, we applied a series of filtering criteria as applicable.

1. For book-based datasets, select books using relevant keywords in the title and then segment them into sentences.
2. Remove sentences that were not part of the main content (e.g., prefaces, tables of contents, copyright statements), as well as sentences containing personal identifiers, such as user tags in social media data.
3. With the exception of domains in which short sentences are typical (e.g., advertising), retain only sentences containing 10–30 words.
4. From the resulting pool of valid sentences, sample up to 2,000 sentences per domain using a fixed random seed (24).

Summary statistics for the dataset are reported in Table 2, and in total, the final dataset used in the acceptability rating experiments comprised 15,162 sentences across the eight domains.

3.3 Prompting Procedure

As described in Section 3.1, the rating is represented as a binary outcome, where 0 indicates an unacceptable sentence and 1 indicates an acceptable one. The prompts used across all experiments were kept consistent for each condition and across models, with the exception of Llama 2 Chat, for

[1] https://dictionary.cambridge.org/dictionary/english/

[2] The source code and other details for dataset reconstruction are available at https://github.com/naist-nlp/caaj

Domain	Form	# Original	# Final
lit	txt (book)	7,957,526	2,000
jou	json (sent)	210,294	2,000
adv	csv (sent)	1,162	1,162
acd	csv (sent)	136,154	2,000
ins	csv (sent)	214,097	2,000
com	txt (book)	233,127	1,000
	csv (sent)	517,401	1,000
onl	csv (sent)	2,782,618	1,500
	csv (sent)	7,398	500
lgl	csv (sent)	17,921	1,020
	txt (book)	124,803	980

Table 2: Statistics of the dataset. *Form* is the original format of source dataset and whether it is book-based or already in sentences. *Original* is the no. of sentences in the source, *Final* is the no. of sentences sampled.

SYSTEM

You are a linguist. You judge whether English sentences are acceptable (natural) for native speakers.
You MUST always respond in valid JSON (no extra text, no markdown) in this shape: {"acceptable":1} or {"acceptable":0}

USER

{task instruction}

Sentence: {sentence}
{additional fields, if any}

Return JSON only with:
- "acceptable":1 if the sentence is acceptable / natural
- "acceptable":0 if the sentence is unacceptable or clearly unnatural
Do not include any explanation or extra fields.

Figure 2: Prompt template

which the system message was merged into the user message. Figure 2 presents the prompt template employed in the experiments where `{sentence}` represents the input sentence to be rated.

Each condition (SO, SK, SKT, and SKD) have different `{task instruction}` as follows:

SO: Judge the natural acceptability based only on the sentence, without using any additional information.

SK: Use the domain keyword as a context hint to judge the natural acceptability of the sentence.

SKT: Use both the domain keyword and its types of work as context to judge the natural acceptability.

SKD: Use both the domain keyword and its definition as context to judge the natural acceptability.

Some conditions may include `{additional fields, if any}` where each field is given by:

- `Keyword: {keyword}`
- `Types: {types}`
- `Definition: {definition}`

The `{keyword}`, `{types}`, and `{definition}` for each domain is as listed in Table 1.

4 Experimental Settings

We performed a set of experiments with multiple LLMs to evaluate English sentences from diverse domains across four distinct contextual conditions.

4.1 Models

We employed six LLMs for inference. Two were proprietary models accessed via API, while the remaining four were open-source models executed using the HuggingFace–vLLM integration (hf_vllm). The models used in the experiments are as follow:

dschat: DeepSeek V3.2 Non-thinking (DeepSeek-AI et al., 2025)

gpt5m: GPT 5 Mini (OpenAI, 2025)

gemma2: Gemma 2 9B Instruct AWQ (Gemma et al., 2024)

llama2: Llama 2 7B Chat AWQ (Touvron et al., 2023)

llama3: Llama 3.1 8B Instruct AWQ (Grattafiori et al., 2024)

qwen2: Qwen 2.5 7B Instruct AWQ (Qwen et al., 2025)

We selected recent lightweight open-source model on the assumption that the binary rating task does not require extensive reasoning. We included medium-sized open models with comparable parameter scales for a more meaningful cross-model comparisons. For all open models, we used 4-bit quantized variants based on Activation-aware Weight Quantization (AWQ), which improves inference efficiency while minimizing accuracy loss.

4.2 Evaluation Methods

Using the acceptability ratings produced by the models under the four conditions, we conduct analyses based on the following quantitative measures.

Aggregated Rating For each domain, compute the percentage of sentences rated as acceptable by each model in each condition. This provides an overview of how the different contextual conditions influence the ratings produced by each model.

Conditional Shift Using SO as the baseline condition, compare the acceptability ratings obtained under SK, SKT, and SKD. This is to examine how, and under what circumstances, additional contextual cues influence the models' judgments. For each sentence, the rating under a contextual condition may remain unacceptable ($0 \rightarrow 0$), become acceptable ($0 \rightarrow 1$), become unacceptable ($1 \rightarrow 0$), or remain acceptable ($1 \rightarrow 1$).

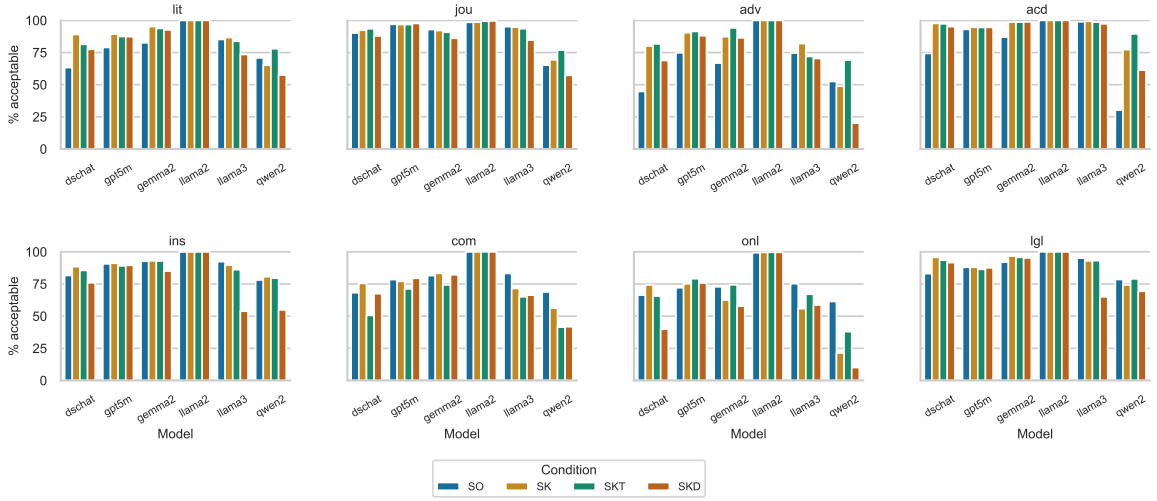

Figure 3: Aggregated acceptability rating for each domain across different conditions

Rating Agreement Treating each model as an independent rater, compute inter-rater agreement using Fleiss' κ. This indicates the extent to which the models exhibit similar rating behaviour, or conversely, whether their judgments diverge in systematic ways. The agreement is computed within each domain and can be evaluated either separately for each condition or aggregated across all conditions.

Rating Stability We introduce a simple scoring scheme to quantify the stability of ratings across conditions. The four conditions are treated as a sequential progression from SO to SK, followed by SKT and SKD. Each sentence–model pair begins with a maximum stability score of 4. For every change in rating observed between two consecutive conditions (for example, SO = 0 and SK = 1), one point is subtracted from the score. Thus, a score of 4 indicates that the rating remains unchanged across all conditions and is therefore highly stable, whereas a score of 1 reflects repeated changes in rating across conditions and corresponds to the lowest level of stability.

5 Experimental Results

5.1 Acceptability Rating Tendency

In general, we hypothesized that providing additional contextual cues would facilitate the models' ability to determine whether a sentence is acceptable. Under our assumption, a sentence is expected to be acceptable within its own domain. Accordingly, we anticipated that the proportion of acceptable sentences should increase relative to the baseline SO condition. Figure 3 illustrates how the proportion of acceptable sentences changes across conditions for each domain. Computation of the confidence intervals, effect sizes, and p-values for each domain is provided in Appendix B.

For several domains, the observed patterns align with this expectation: the SO condition yields lower acceptance rates, while the contextual conditions produce higher rates. Comparing the three non-baseline conditions, we find that providing exemplar-type information (SKT) is generally more effective in increasing domain-consistent acceptability, whereas supplying definitional information (SKD) does not yield comparable gains. This suggests that concrete illustrative cues are more useful than abstract elaboration to guide the model.

Across models, however, notable divergence is observed. In particular, llama2 and qwen2 show tendencies that differ from the other models. The deviation exhibited by llama2 is plausibly attributable to the fact that it is not instruction-tuned, which may limit its sensitivity to contextual cues in the prompt. By contrast, the comparatively low acceptance rates produced by qwen2 appear to reflect conservative internal behaviour: although its rating trends are directionally similar to those of some other models, its overall threshold for deeming a sentence acceptable is substantially stricter.

To account for variability in acceptability ratings due to prompt sensitivity, we conducted a validation study using 200 randomly selected sentences from each of the eight domains. The evaluation includes two proprietary models (dschat and gpt5m)

dschat			
Dom	**SO → SK**	**SO → SKT**	**SO → SKD**
lit	+24.0 (±6.4)	+16.0 (±11.2)	+11.5 (±11.0)
jou	+5.5 (±3.7)	+6.5 (±3.4)	+1.5 (±6.8)
adv	+36.0 (±12.3)	+37.5 (±10.0)	+23.5 (±14.4)
acd	+24.5 (±4.2)	+24.0 (±4.5)	+22.0 (±4.8)
ins	+7.5 (±4.7)	+3.5 (±5.4)	-5.5 (±9.6)
com	+10.0 (±3.8)	-18.0 (±16.9)	+3.0 (±12.2)
onl	+7.0 (±6.9)	-3.5 (±8.4)	-27.0 (±16.3)
lgl	+16.5 (±4.1)	+15.0 (±5.4)	+12.5 (±8.7)
gpt5m			
Dom	**SO → SK**	**SO → SKT**	**SO → SKD**
lit	+11.5 (±3.4)	+8.5 (±2.9)	+8.0 (±4.0)
jou	-0.5 (±2.5)	+0.0 (±0.9)	+2.0 (±4.1)
adv	+15.0 (±3.9)	+15.0 (±4.3)	+10.5 (±3.3)
acd	+4.0 (±1.4)	+5.0 (±0.6)	+6.5 (±3.0)
ins	+0.5 (±1.9)	+0.0 (±5.4)	+0.0 (±2.9)
com	-4.5 (±0.8)	-11.0 (±4.4)	-0.5 (±0.9)
onl	-3.0 (±2.1)	+6.0 (±3.4)	-1.0 (±3.3)
lgl	+0.0 (±1.8)	-1.0 (±4.9)	-1.0 (±0.2)
llama3			
Dom	**SO → SK**	**SO → SKT**	**SO → SKD**
lit	+0.0 (±7.8)	-3.5 (±4.0)	-13.5 (±3.5)
jou	+2.5 (±4.3)	+1.5 (±2.6)	-5.5 (±2.6)
adv	+11.0 (±11.2)	-4.0 (±3.8)	-5.0 (±0.5)
acd	+0.5 (±2.7)	-0.5 (±1.9)	-3.0 (±2.2)
ins	-1.5 (±2.8)	-7.5 (±1.4)	-36.5 (±12.0)
com	-8.5 (±4.4)	-18.5 (±4.2)	-16.0 (±5.2)
onl	-19.5 (±4.4)	-11.0 (±3.0)	-18.5 (±4.1)
lgl	+0.0 (±3.0)	+0.0 (±2.2)	-28.5 (±4.0)
qwen2			
Dom	**SO → SK**	**SO → SKT**	**SO → SKD**
lit	-5.5 (±1.7)	+5.0 (±2.9)	-13.5 (±1.5)
jou	+8.5 (±5.1)	+18.0 (±7.7)	-3.5 (±6.2)
adv	+0.5 (±3.4)	+16.5 (±2.5)	-28.0 (±4.1)
acd	+49.0 (±10.0)	+60.5 (±15.4)	+34.5 (±11.9)
ins	+5.0 (±1.2)	+4.0 (±1.0)	-19.5 (±5.0)
com	-11.5 (±2.9)	-25.5 (±0.5)	-27.0 (±3.4)
onl	-41.0 (±3.5)	-25.0 (±1.6)	-53.0 (±2.2)
lgl	-6.0 (±2.5)	+1.5 (±3.8)	-9.5 (±3.2)

Table 3: Acceptability ratings across prompt variations

and two open-source models (llama3 and qwen2). The prompt variations comprise: 1) Altering the position of the sentence relative to the contextual cues for SK, SKT, and SKD, 2) Randomizing the order of cues in SKT and SKD, 3) Using neutral instead of directive task instructions without substantial paraphrasing, 4) Paraphrasing the task instructions while maintaining a directive tone, and 5) Modifying the rating format from 0/1 to a/b. The corresponding instructions for variants 3 and 4 are provided in Appendix C.

Using aggregated ratings, we compute the change in the proportion of acceptable sentences for each condition relative to SO. Table 3 presents the results with standard deviations reported in parentheses. The results indicate that the overall patterns are generally robust to these prompt variations with only few domains deviating.

5.2 Shifts in Acceptability Rating

Across all domains, a substantial proportion of sentences were judged acceptable by most models in the baseline condition, and these sentences generally remained acceptable under the other conditions. Figure 6 in Appendix D summarizes the distribution of rating transitions relative to the SO baseline.

To focus on cases in which contextual cues changed the decision outcome, we specifically look at sentences whose ratings change, i.e., those that become unacceptable ($1 \rightarrow 0$) or become acceptable ($0 \rightarrow 1$). This allows us to assess whether particular cues tend to make a model more lenient or more stringent in its judgments. To quantify this tendency, we compute the difference between the proportion of $0 \rightarrow 1$ transitions and the proportion of $1 \rightarrow 0$ transitions. These proportions are calculated with respect to the subset of sentences that exhibit rating changes, rather than the full dataset. The resulting values are shown in Figure 4.

Greater values indicate that majority of changed ratings shifts towards acceptance ($0 \rightarrow 1$), implying that the model becomes more lenient under the contextual condition. Smaller values indicate the opposite pattern, that contextual cues lead the model to reject sentences it previously accepted, reflecting a more stringent judgment. The results reveal domain-specific tendencies: several models become more lenient in the literature, advertisement, academic, and legal domains, where contextual cues appear to support reinterpretation of sentences as appropriate. These results also reinforce the divergent behaviour of llama2 and qwen2. Llama2 either exhibits no rating change or shifts strongly toward leniency, whereas qwen2 tends to become more stringent relative to the baseline.

5.3 Agreement of Models

The preceding analyses examined acceptability ratings at an aggregate level. We now turn to the level of individual sentences to assess the extent to which the models (i.e., raters) agree in their judgments. For each condition, we group sentences within a domain and compute inter-rater agreement using Fleiss' κ. This metric captures whether the models tend to converge on the same rating for a given sentence or whether their judgments diverge. Table 4 reports the agreement scores for each domain under all conditions and the mean agreement values.

The results show that agreement among models is low. The lowest mean agreement within a do-

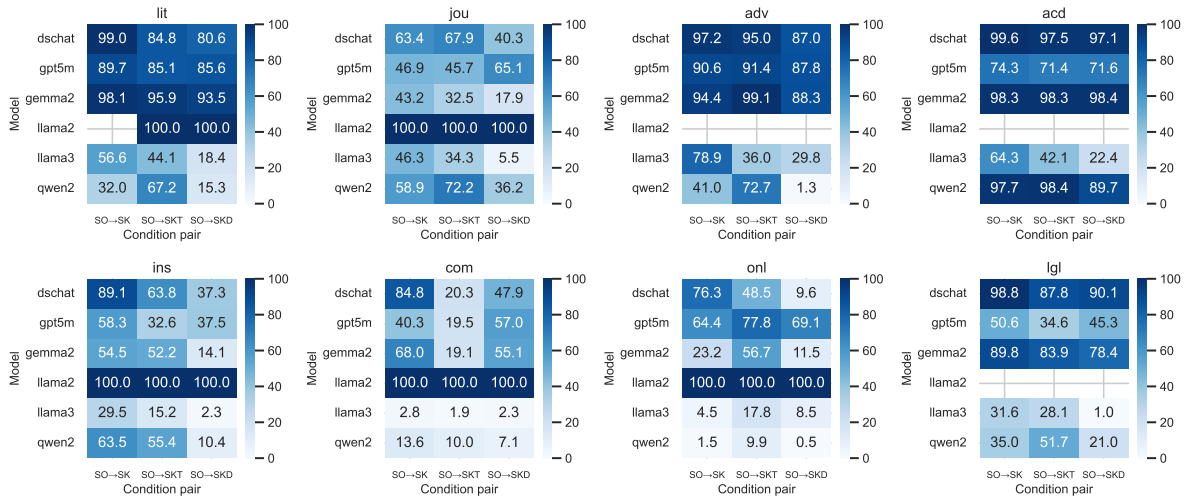

Figure 4: Nature of the changes of acceptability rating in each domain

Domain	SO	SK	SKT	SKD	*MeanD*
lit	.317	.228	**.300**	**.245**	.272
jou	.131	.175	.227	.200	.183
adv	.339	.220	.217	.144	.230
acd	.013	.074	.177	.054	.080
ins	.302	.301	.285	.224	.278
com	.355	**.312**	.260	.234	**.290**
onl	**.364**	.175	.205	.106	.213
lgl	.274	.198	.262	.202	.234
MeanC	**.262**	.210	.242	.176	—

Table 4: Fleiss' κ per domain and across conditions. Highest κ value is in **bold**, while lowest is in underline. *MeanD* column indicates mean score of each domain, *MeanC* row indicates mean score of each condition.

main (MeanD) is observed in the academic domain, while the highest occurs in the communication domain. Referring to the academic panel in Figure 3, we note that qwen2 exhibits behaviour that diverges substantially from the other models, which likely contributes to the reduced agreement scores.

When examining mean agreement within a condition (MeanC), the highest agreement occurs under the baseline SO condition. This suggests that, in the absence of contextual cues, as the models rely more directly on prior training signals, they tend to converge in their judgments. Then, as contextual cues are introduced, the models increasingly diverge, indicating that different models appear to incorporate contextual cues in distinct ways.

5.4 Stability of Ratings

Using the stability scoring scheme described earlier, we examine how each model behaves across domains. Figure 5 presents the distribution of stability scores for each domain, faceted by model. Brighter polygons correspond to higher stability scores, with a score of 4 (i.e., no change in rating across conditions) represented by the brightest shading and a score of 1 (i.e., frequent rating changes across conditions) represented by the darkest.

The patterns indicate that gpt5m and gemma2 show a broadly similar behaviour: lower stability scores are concentrated near the centre of the plot, suggesting that relatively few sentences undergo repeated rating changes. A similar distribution is observed for dschat and llama3, although in these models the proportion of highly stable ratings is smaller, indicating a greater degree of contextual sensitivity. Within this pair, stability is comparatively well preserved in the journalistic and legal domains for dschat, whereas llama3 shows greater stability in the academic domain.

In contrast, the most divergent stability profiles are as expected observed for llama2 and qwen2. Qwen2 exhibits high stability in the legal domain but considerably lower stability in the academic, advertisement, and online-media domains. These differing stability patterns suggest that models vary in how readily their judgments can be influenced by contextual cues. In domains where stability is high, contextual cues exert relatively little effect on the rating outcome, whereas lower stability indicates that the model is more easily "nudged" by the additional contextual prompts.

6 Analysis and Discussion

A closer examination of domain characteristics suggests that domains containing a higher proportion of ungrammatical sentences, uncommon lexi-

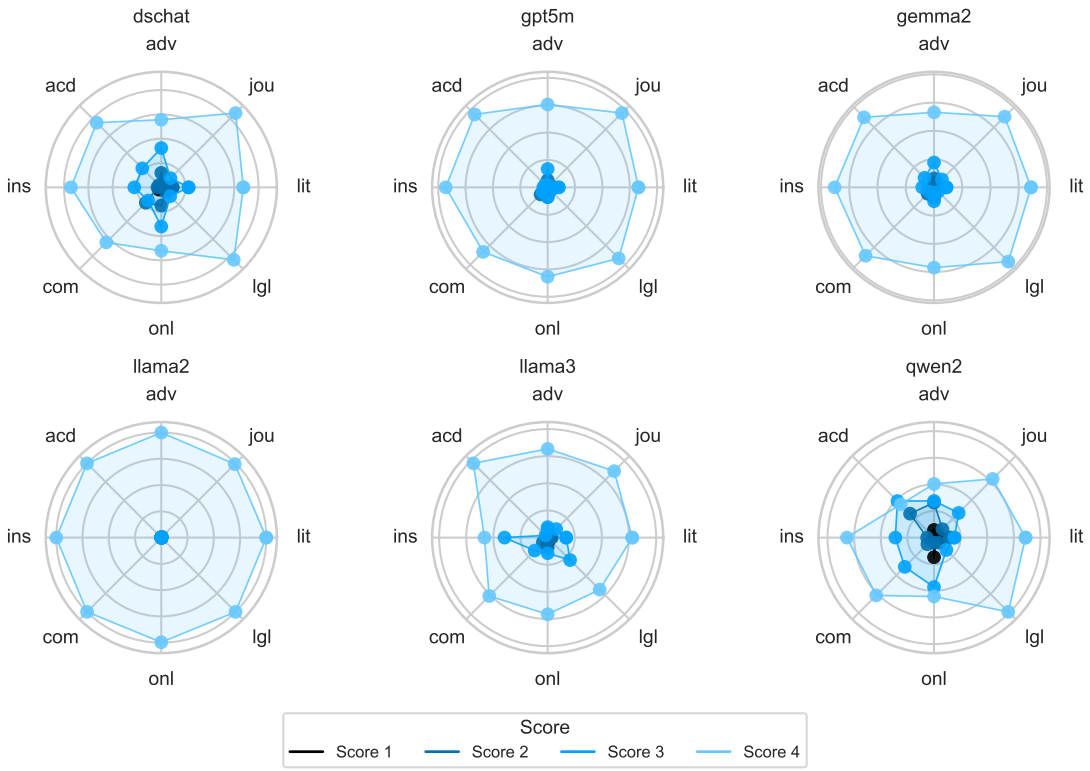

Figure 5: Stability scores of each model on different domains

cal items, or specialized terminology benefit more from contextual cues. Significant increases in acceptability are observed in the literature, advertisement, and academic domains. Literary texts often employ rare or stylistically marked vocabulary, advertisement language frequently deviates from canonical grammatical structure, and academic writing relies heavily on technical terminology. In such cases, domain cues appear to help models interpret sentences in context rather than evaluating them as isolated strings. To support this claim, representative examples of sentences from each domain are presented in Appendix E.

Next, a deeper look into domains with lower inter-model agreement suggests that sentences containing domain-specific or technical terminology tend to elicit less consensus among models. Interestingly, we initially expected a similar pattern in the advertisement domain, as such texts are often ungrammatical or fragmentary. However, the relatively higher agreement observed in this domain may reflect the models' extensive exposure to such text types, resulting in more consistent ratings.

Focusing only on sentences with a stability score below 4–that is, sentences whose ratings change at least once–we examine which specific contextual elements trigger changes in the models' acceptability judgments across domains and lead to stable outcomes. In particular, we analyse sentences with a stability score of 3 that follow the characteristic patterns. The patterns 0111 and 1000 indicate a rating change that occurs when the domain keyword is introduced; we refer to this pattern as **SK**. The patterns 0011 and 1100 indicate a rating change that occurs when the keyword is introduced together with additional contextual cue (in the sequence that we defined: Types); we refer to this pattern as **SKX**. Table 5 reports the percentage of sentences exhibiting each pattern, along with the mean percentage across domains.

The results in the table indicate, for each model, which contextual element is more influential when a rating changes and subsequently stabilizes—whether the change is triggered by the keyword alone or by the keyword in combination with additional information. In most cases, the rating change occurs at the point where the domain keyword is introduced, suggesting that, even in the absence of further elaboration, this cue is sufficient to prompt a reinterpretation of the sentence's acceptability. The table also makes it possible to identify, on a per-domain basis, which type of contextual cue is more effective in prompting a model to revise its judgment and settle into a stable rating.

Domain	dschat		gpt5m		gemma2		llama2		llama3		qwen2	
	SK	SKX	SK	SKX	SK	SKX	SK	SKX	SK	SKX	SK	SKX
lit	54.7	9.5	47.1	6.2	73.8	1.4	0.0	**100.0**	30.2	10.4	27.9	2.7
jou	45.2	9.2	25.0	7.5	38.1	**8.7**	9.5	71.4	18.7	11.2	44.6	2.4
adv	65.1	1.6	**58.2**	5.3	51.5	7.9	—	—	25.5	**22.2**	11.4	6.1
acd	**86.3**	1.4	31.3	7.0	**94.4**	0.8	—	—	34.4	17.2	**54.7**	1.2
ins	26.5	19.5	23.4	13.1	21.6	5.0	**100.0**	0.0	11.0	9.1	15.9	10.0
com	13.9	**24.8**	15.1	9.0	21.3	**8.7**	**100.0**	0.0	**40.8**	14.4	40.9	**31.0**
onl	11.8	18.0	27.3	9.6	20.1	2.1	75.0	12.5	26.4	5.6	45.2	2.2
lgl	58.6	4.8	18.3	**14.0**	65.7	3.6	—	—	8.7	2.2	44.4	2.9
Mean	45.3	11.1	30.7	9.0	48.3	4.8	56.9	36.8	24.5	11.5	35.6	7.3

Table 5: Percentage of rating change condition. Highest value is in **bold**. Lowest value is in underline

Using the same randomly sampled dataset as that in the prompt variation validation, we examine acceptability ratings across several open-source models within the same family. The results indicate broadly comparable acceptance rates across models, supporting the robustness of the observed patterns. A detailed breakdown of acceptance rates for each condition is provided in Appendix F. For this validation, Gemma 3 models are evaluated in full precision rather than in quantized form. Consistent with earlier observations, Qwen 2 exhibits distinct behaviour relative to the other models, reflecting differences in its architecture and training.

The aforementioned findings support the view that acceptability is not evaluated as a purely sentence-internal property, but depends on how a sentence is interpreted within a given context. The observed effects of domain cues indicate that even minimal contextual information can influence model judgments. At the same time, the variation across models suggests that this sensitivity is not consistently represented, but depends on model-specific characteristics. From a linguistic perspective, this aligns with the distinction between grammaticality and acceptability, while also indicating that models capture this distinction only partially.

More broadly, the results clarify what is being measured when LLMs are used as acceptability raters. The judgments produced reflect an interaction between learned distributional patterns and the contextual information provided at inference time, rather than a stable notion of grammatical competence. This has implications for how such judgments are interpreted in downstream applications, where outcomes may vary depending on how context is specified. It also suggests limits to generalization—an issue further discussed in the limitations of this study: evaluations conducted under simplified or controlled conditions may not fully reflect model behaviour in more complex or less well-defined contexts. Further investigation is therefore needed to better understand how contextual factors shape model judgments across a wider range of settings.

7 Conclusion

This study investigates how contextual cues affect LLM acceptability judgments across multiple domains. The results show that context–in the form of domain framing rather than surrounding discourse–can meaningfully influence model behaviour, but its effects are uneven across models and domains. In particular, models tend to benefit more from concrete, example-oriented cues than from abstract explanatory information, and some domains–such as, literature, advertisement, and academic writing–exhibit greater sensitivity to contextual support, likely due to atypical lexical, structural, or stylistic properties. At the same time, the models do not converge on a unified interpretation strategy: their acceptance thresholds and responses to context diverge, reflecting underlying differences in how acceptability is internally represented.

Beyond changes in rating outcomes, analyses of agreement and stability reveal that models frequently disagree at the sentence level and that their judgments become less aligned when contextual (domain) cues are introduced, suggesting that the cues are incorporated in model-specific ways. Stability patterns further indicate when ratings shift and which kinds of cues are most likely to trigger reinterpretation. Taken together, these findings provide a systematic account of how domain-framing context shapes model-based acceptability judgments, highlighting both the potential and the limitations of contextual prompting. This work thus contributes empirical evidence for understanding context-sensitive model behaviour and offers a basis for developing more robust approaches to evaluating acceptability in LLMs.

Limitations

This study focuses on a specific set of large language models and model configurations. As such, the findings should not be taken as exhaustive or universally generalizable, as different architectures, instruction-tuning strategies, model sizes, or future versions may exhibit different behaviours. In addition, the open-source models were evaluated in quantized form, which may introduce minor deviations relative to full-precision inference. Further evaluation across a broader range of models and configurations would be needed to assess the extent to which these findings generalize.

The experiments are conducted exclusively on English data. Therefore, the results reflect model behaviour within a single language and its associated conventions, and may not directly transfer to languages with different morphological, syntactic, or discourse properties. Extending the analysis to multilingual settings would be necessary to determine whether similar effects hold more broadly.

The task is operationalized using a binary acceptability rating and a fixed prompt design. While this enables controlled comparison across models and conditions, binary ratings may obscure gradience in borderline or ambiguous cases, and alternative task formulations (e.g., scalar judgments or justification-based prompts) may lead to different outcomes. Future work could explore how different task designs affect model sensitivity to contexts.

The study further assumes that sentences are expected to be acceptable within their own domain, which provides a useful analytical baseline but does not capture all pragmatic or stylistic uses of language. Some sentences may be intentionally deviant, creative, or context-dependent in ways that are not fully reflected by domain labels. In addition, the dataset is constructed from publicly available written sources and filtered using heuristic criteria, which may introduce sampling bias and does not represent spoken or conversational language. Broader data sources and more fine-grained domain distinctions could help refine these assumptions.

Context in this study is represented through controlled cues such as keywords, text types, and definitions. While this design allows us to isolate specific contextual effects, it does not capture richer forms of context. Investigating the more complex forms of context would provide a more complete picture of how acceptability judgments are shaped in practice.

Finally, the analyses characterize observable model behaviour, including agreement and rating stability, rather than underlying reasoning processes or linguistic competence. The study does not establish alignment with human judgments, and the extent to which model behaviour reflects human acceptability intuitions remains an open question. Incorporating human evaluation would be an important next step in assessing this relationship.

Ethical Considerations

The datasets used in this study are drawn from publicly available sources across multiple domains. Their use follows the respective terms and licenses under which they were released. Although filtering procedures were applied, including the removal of metadata and personal identifiers where applicable, the data may still contain informal, biased, or otherwise sensitive language inherent to naturally occurring text. These characteristics reflect the original sources and may influence model behaviour.

The study focuses on analysing acceptability judgments produced by large language models and does not involve deploying systems in real-world applications. The findings indicate that such judgments are sensitive to contextual framing, which may affect how outputs are interpreted in downstream settings. In applications such as automated evaluation or content moderation, this sensitivity suggests that model judgments should not be treated as fixed or definitive indicators, but as context-dependent outputs that require careful interpretation.

In the preparation of the manuscript, ChatGPT was used to assist with language polishing and editing. However, all experimental design, implementation, analysis, and interpretive decisions were carried out by the authors. The authors remain fully responsible for the content, accuracy, and contributions of the paper.

Acknowledgments

The authors thank Hidetaka Kamigaito for valuable feedback and discussions. We also thank the anonymous reviewers and the action editor for their helpful comments. This work made use of API credits and computational resources provided by the Natural Language Processing Laboratory of Nara Institute of Science and Technology, Japan. This work has been partially supported by JSPS KAKENHI Grant Numbers 25K24369 and 26K21312.

References

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

David Beauchemin and Richard Khoury. 2025. QFrCoLA: a Quebec-French corpus of linguistic acceptability judgments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 119–130, Suzhou, China. Association for Computational Linguistics.

Jean-Philippe Bernardy, Shalom Lappin, and Jey Han Lau. 2018. The influence of context on sentence acceptability judgements. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 456–461, Melbourne, Australia. Association for Computational Linguistics.

Yuri Bizzoni and Shalom Lappin. 2018. Predicting human metaphor paraphrase judgments with deep neural networks. In *Proceedings of the Workshop on Figurative Language Processing*, pages 45–55, New Orleans, Louisiana. Association for Computational Linguistics.

Vijay Daultani, Héctor Javier Vázquez Martínez, and Naoaki Okazaki. 2024. Acceptability evaluation of naturally written sentences. *Journal of Information Processing*, 32:652–666.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Venkatesh E, Kaushal Maurya, Deepak Kumar, and Maunendra Sankar Desarkar. 2023. DivHSK: Diverse headline generation using self-attention based keyword selection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1879–1891, Toronto, Canada. Association for Computational Linguistics.

Gisbert Fanselow. 2021. Acceptability, grammar, and processing. In GrantEditor Goodall, editor, *The Cambridge Handbook of Experimental Syntax*, Cambridge Handbooks in Language and Linguistics, page 118–153. Cambridge University Press.

Elaine J. Francis. 2021. *Gradient Acceptability and Linguistic Theory*. Oxford University Press.

Team Gemma, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Team Gemma, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. *NeurIPS*.

Benedetta Iavarone, Dominique Brunato, and Felice Dell'Orletta. 2021. Sentence complexity in context. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 186–199, Online. Association for Computational Linguistics.

Yusuke Ide, Yuto Nishida, Justin Vasselli, Miyu Oba, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2025. How to make the most of LLMs' grammatical knowledge for acceptability judgments. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7416–7432, Albuquerque, New Mexico. Association for Computational Linguistics.

Tom S Juzek. 2015. *Acceptability judgement tasks and grammatical theory*. Ph.D. thesis, University of Oxford.

Tom S Juzek. 2024. The syntactic acceptability dataset (preview): A resource for machine learning and linguistic analysis of English. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16113–16120, Torino, Italia. ELRA and ICCL.

Shibamouli Lahiri. 2014. Complexity of word collocation networks: A preliminary structural analysis. In

Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 96–105, Gothenburg, Sweden. Association for Computational Linguistics.

Steven Langsford, Amy Perfors, Andrew T. Hendrickson, Lauren A. Kennedy, and Danielle J. Navarro. 2018. Quantifying sentence acceptability measures: Reliability, bias, and variability. *Glossa: a journal of general linguistics*, 3(1)(37).

Jey Han Lau, Carlos Armendariz, Shalom Lappin, Matthew Purver, and Chang Shu. 2020. How furiously can colorless green ideas sleep? sentence acceptability in context. *Transactions of the Association for Computational Linguistics*, 8:296–310.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive science*, 41(5):1202–1241.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.

Vladislav Mikhailov, Tatiana Shamardina, Max Ryabinin, Alena Pestova, Ivan Smurov, and Ekaterina Artemova. 2022. RuCoLA: Russian corpus of linguistic acceptability. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5207–5227, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rishabh Misra. 2022. News category dataset. *Preprint*, arXiv:2209.11429.

Rishabh Misra and Jigyasa Grover. 2021. *Sculpting Data for ML: The first act of Machine Learning*. Independently published.

James Myers. 2017. Acceptability judgments.

OpenAI. 2025. Gpt-5 system card.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Karolina Rudnicka. 2025. The language of ai tools as idiolects -thus comparable to other idiolects.

Koustuv Sinha, Jon Gauthier, Aaron Mueller, Kanishka Misra, Keren Fuentes, Roger Levy, and Adina Williams. 2023. Language model acceptability judgements are not always robust to context. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6043–6063, Toronto, Canada. Association for Computational Linguistics.

Taiga Someya, Yushi Sugimoto, and Yohei Oseki. 2024. JCoLA: Japanese corpus of linguistic acceptability. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9477–9488, Torino, Italia. ELRA and ICCL.

Jon Sprouse, editor. 2023. *The Oxford Handbook of Experimental Syntax*, chapter Acceptability judgment methods. Oxford University Press.

Jon Sprouse and Diogo Almeida. 2017. Design sensitivity and statistical power in acceptability judgment experiments. *Glossa: a journal of general linguistics*, 2(1)(14):1–32.

Jon Sprouse and Carson Schütze. 2019. Grammar and the use of data. In Bas Aarts, Jill Bowie, and Gergana Popova, editors, *The Oxford Handbook of English Grammar*. Oxford University Press.

Jon Sprouse, Carson T. Schütze, and Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001–2010. *Lingua*, 134:219–248.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Héctor Javier Vázquez Martínez. 2021. The acceptability delta criterion: Testing knowledge of language using the gradience of sentence acceptability. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 479–495, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Peinan Zhang, Yusuke Sakai, Masato Mita, Hiroki Ouchi, and Taro Watanabe. 2025. AdTEC: A unified benchmark for evaluating text quality in search engine advertising. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7672–7691, Albuquerque, New Mexico. Association for Computational Linguistics.

A Public Dataset Sources

The dataset used in the experiments was constructed from the specified sources.

Literary Works :
(Lahiri, 2014)
https://shibamoulilahiri.github.io/gutenberg_dataset.html

Journalistic :
(Misra and Grover, 2021; Misra, 2022)
https://www.kaggle.com/datasets/rmisra/news-category-dataset

Advertisements :
https://www.kaggle.com/datasets/chaibapat/slogan-dataset

Academic/Scientific :
https://www.kaggle.com/datasets/sumitm004/arxiv-scientific-research-papers-dataset

Instructional/Technical :
https://www.kaggle.com/datasets/varunucl/wikihow-summarization

Functional/Everyday :
(Lahiri, 2014)
https://shibamoulilahiri.github.io/gutenberg_dataset.html
https://www.kaggle.com/datasets/wcukierski/enron-email-dataset

Digital/Online :
https://www.kaggle.com/datasets/glushko/seth-godins-blogs-dataset
https://www.kaggle.com/datasets/thoughtvector/customer-support-on-twitter

Legal/Official Documents :
(Hendrycks et al., 2021)
https://www.kaggle.com/datasets/konradb/atticus-open-contract-dataset-aok-beta
https://www.kaggle.com/datasets/shivamb/legal-citation-text-classification

B Computation of Confidence Intervals (CI), Effect Sizes, and P-values

Using our existing results, we have computed the 95% confidence intervals (CI) for acceptability rating in SO condition, effect sizes (Δ) for all other conditions with respect to SO, and p-values (p) with False Discovery Rate (FDR) adjustment using the Benjamini-Hochberg (BH) procedure. Each of Tables 6 to 13 represents a domain. The Δ values for each condition are represented in percentage-point.

C Prompt Variations

Following are the variants of the original directive instructions used in the validation study.

Neutral Instructions

SO: Judge the natural acceptability based only on the sentence, without using any additional information.

SK: Judge the natural acceptability of the sentence. You may consider the domain keyword as a context hint.

SKT: Judge the natural acceptability of the sentence. You may consider the domain keyword and its types of work as context hint.

SKD: Judge the natural acceptability of the sentence. You may consider the domain keyword and its definition as a context hint.

Paraphrased Directive Instructions 1

SO: Without using any additional information, consider the linguistic integrity of the sentence.

SK: Use the domain keyword as a context hint to consider the linguistic integrity of the sentence.

SKT: Use both the domain keyword and its types of work as context to consider the linguistic integrity of the sentence.

SKD: Use both the domain keyword and its definition as context to consider the linguistic integrity of the sentence.

Paraphrased Directive Instructions 2

SO: Without using any additional information, judge the correctness of the sentence by the strictest standards.

SK: Use the domain keyword as a context hint to judge the correctness of the sentence by the strictest standards.

SKT: Use both the domain keyword and its types of work as context to judge the correctness of the sentence by the strictest standards.

SKD: Use both the domain keyword and its definition as context to judge the correctness of the sentence by the strictest standards.

D Overall Change of Acceptability Rating

Figure 6 summarizes the distribution of rating transitions in each domain for each model relative to the SO baseline.

E Example of Sentences from Each Domain

Following are several examples of sentences in the dataset for each domain. Each of these sentences is initially judged as unacceptable by several models (at SO) and then, as we introduce contextual domain cues (at SK, SKT, or SKD), it becomes acceptable. Note that not all rating changes happen at the same condition.

Literature

- Till as I gaze with staring eyes,
- Or why so long (in life if long can be) Lent Heaven a parent to the poor and me?

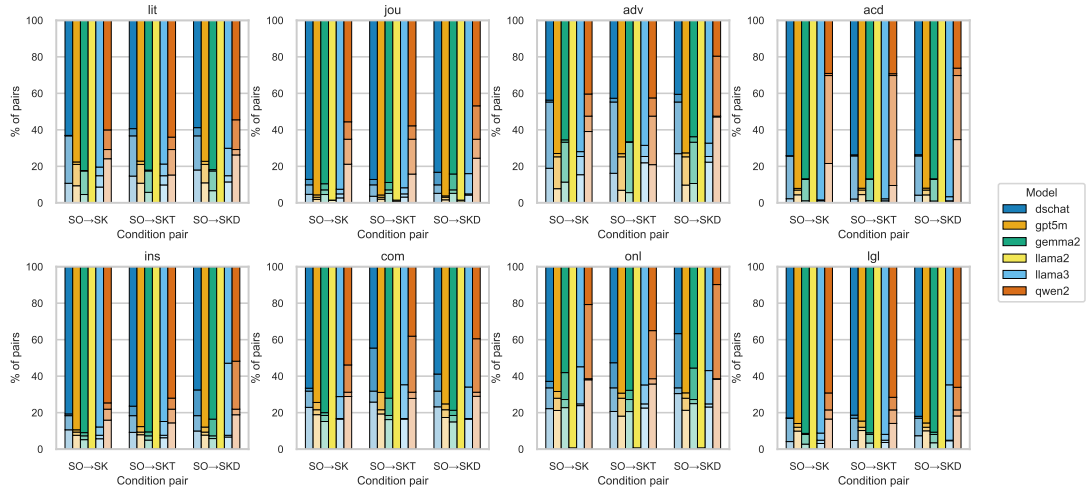

Figure 6: Change of acceptability ratings in each domain, across different conditions, in comparison to the baseline. From top to bottom (darker to lighter colour), each part corresponds to one of the following changes of rating respectively: remains acceptable ($1 \rightarrow 1$), becomes unacceptable ($1 \rightarrow 0$), becomes acceptable ($0 \rightarrow 1$), remains unacceptable ($0 \rightarrow 0$).

Journalistic

- Chewing, And Choking, On False (Nutritional) Equivalence
- A REAL PIZZA WORK: Parolee Trapped In Pizza Vent For 11 Hours

Advertisement

- Drink Fanta, stay Bamboocha.
- Discover your baby's world.

Academic

- Superclustering by finding statistically significant separable groups of optimal gaussian
- Evaluation the efficiency of artificial bee colony and the firefly algorithm in solving the continuous optimization problem

Instructional

- Drag an image from from a Rox-Filer window into the resulting window to change the background.
- Slide "Rest Finger to Open" to the on or green position.

Communication

- Some of those who were essentially in harmony with his views preceded, and many followed him.
- Our group is working on PortRac, a database that tracks the assets activities, and I need information for first quarter 2001.

Online Media

- On the live chat talking to someone right now:) was not impressed that last time the driver visited he let himself into my flat!
- We regret your experience. Could you please elaborate your exact concern, so that we can assist you in a better way. Pooja OPPO Care

Legal

- Logan's also shall have the exclusive right to promote its Logan's trademark on the side of the helmet, as shown on Exhibit B to this Agreement.
- Grace or to such other address as the parties hereto may specify, in writing, from time to time.

F Acceptability Rating Across Open-Source Models

For the following Tables 14 to 17, each table represents a condition. To distinguish the models, the number of parameters are also written after the model name. The values in parentheses indicate differences relative to the originally selected model in *italic*. Aside from the originally selected model, other variants used are:

gemma3_12b: Gemma 3 27B Instruct (Gemma et al., 2025)

gemma3_27b: Gemma 3 27B Instruct

llama3_70b: Llama 3.1 70B Instruct AWQ

qwen2_72b: Qwen 2.5 72B Instruct AWQ

Model	SO (CI)	SK		SKT		SKD	
		Δ	p	Δ	p	Δ	p
dschat	0.633 (0.612–0.654)	+25.75	< .001	+18.10	< .001	+14.20	< .001
gpt5m	0.789 (0.771–0.807)	+10.35	< .001	+8.50	< .001	+8.40	< .001
gemma2	0.826 (0.808–0.842)	+12.65	< .001	+11.20	< .001	+10.05	< .001
llama2	1.000 (0.997–1.000)	+0.00	< .001	+0.05	1.000	+0.05	1.000
llama3	0.852 (0.835–0.866)	+1.45	.073	-1.35	.105	-11.65	< .001
qwen2	0.708 (0.688–0.728)	-5.70	< .001	+7.15	< .001	-13.35	< .001

Table 6: CI, Δ, and p for **literature (lit)** domain

Model	SO (CI)	SK		SKT		SKD	
		Δ	p	Δ	p	Δ	p
dschat	0.901 (0.888–0.914)	+2.20	.001	+3.30	< .001	-2.25	.005
gpt5m	0.970 (0.961–0.976)	-0.15	.827	-0.20	.713	+0.65	.083
gemma2	0.929 (0.917–0.939)	-0.80	.199	-2.10	< .001	-6.80	< .001
llama2	0.985 (0.979–0.989)	+0.10	.539	+0.85	< .001	+1.05	< .001
llama3	0.951 (0.941–0.960)	-0.35	.592	-1.60	.003	-10.45	< .001
qwen2	0.651 (0.630–0.672)	+4.15	< .001	+11.75	< .001	-7.90	< .001

Table 7: CI, Δ, and p for **journalistic (jou)** domain

Model	SO (CI)	SK		SKT		SKD	
		Δ	p	Δ	p	Δ	p
dschat	0.448 (0.419–0.476)	+35.28	< .001	+37.01	< .001	+24.10	< .001
gpt5m	0.748 (0.722–0.772)	+15.66	< .001	+16.61	< .001	+13.34	< .001
gemma2	0.668 (0.640–0.694)	+20.57	< .001	+27.37	< .001	+19.62	< .001
llama2	1.000 (0.997–1.000)	+0.00	< .001	+0.00	< .001	+0.00	< .001
llama3	0.746 (0.720–0.770)	+7.31	< .001	-2.67	.006	-4.22	< .001
qwen2	0.525 (0.496–0.554)	-3.70	.009	+16.61	< .001	-32.44	< .001

Table 8: CI, Δ, and p for **advertisement (adv)** domain

Model	SO (CI)	SK		SKT		SKD	
		Δ	p	Δ	p	Δ	p
dschat	0.743 (0.723–0.762)	+23.35	< .001	+23.00	< .001	+20.85	< .001
gpt5m	0.930 (0.918–0.940)	+1.70	< .001	+1.50	< .001	+1.60	< .001
gemma2	0.871 (0.855–0.885)	+11.65	< .001	+11.60	< .001	+11.75	< .001
llama2	1.000 (0.998–1.000)	+0.00	< .001	+0.00	< .001	+0.00	< .001
llama3	0.989 (0.983–0.993)	+0.40	.219	-0.30	.481	-1.60	< .001
qwen2	0.302 (0.283–0.323)	+47.05	< .001	+59.25	< .001	+31.05	< .001

Table 9: CI, Δ, and p for **academic (acd)** domain

Model	SO (CI)	SK		SKT		SKD	
		Δ	p	Δ	p	Δ	p
dschat	0.817 (0.799–0.833)	+6.80	< .001	+3.95	< .001	-5.70	< .001
gpt5m	0.906 (0.893–0.918)	+0.50	.287	-1.55	.002	-1.10	.032
gemma2	0.927 (0.915–0.938)	+0.35	.552	+0.20	.808	-7.65	< .001
llama2	0.999 (0.996–1.000)	+0.05	1.000	+0.05	1.000	+0.05	1.000
llama3	0.924 (0.912–0.935)	-2.65	< .001	-6.20	< .001	-38.55	< .001
qwen2	0.781 (0.762–0.799)	+2.55	< .001	+1.45	.108	-23.25	< .001

Table 10: CI, Δ, and p for **instructional (ins)** domain

Model	SO (CI)	SK		SKT		SKD	
		Δ	p	Δ	p	Δ	p
dschat	0.682 (0.661–0.702)	+7.30	< .001	-17.60	< .001	-0.75	.522
gpt5m	0.784 (0.765–0.801)	-1.30	.039	-7.20	< .001	+1.05	.122
gemma2	0.816 (0.798–0.832)	+1.75	< .001	-7.25	< .001	+0.65	.333
llama2	1.000 (0.997–1.000)	+0.05	1.000	+0.05	1.000	+0.05	1.000
llama3	0.832 (0.815–0.848)	-11.65	< .001	-18.15	< .001	-16.85	< .001
qwen2	0.688 (0.667–0.708)	-12.60	< .001	-27.30	< .001	-27.05	< .001

Table 11: CI, Δ, and p for **communication (com)** domain

Model	SO (CI)	SK		SKT		SKD	
		Δ	p	Δ	p	Δ	p
dschat	0.664 (0.643–0.684)	+7.85	< .001	-0.80	.572	-26.55	< .001
gpt5m	0.721 (0.701–0.740)	+3.00	< .001	+7.00	< .001	+3.65	< .001
gemma2	0.728 (0.708–0.747)	-10.30	< .001	+1.55	.062	-15.00	< .001
llama2	0.993 (0.988–0.996)	+0.30	.052	+0.40	.017	+0.35	.030
llama3	0.752 (0.733–0.770)	-19.40	< .001	-8.15	< .001	-16.55	< .001
qwen2	0.614 (0.592–0.635)	-40.05	< .001	-23.45	< .001	-51.35	< .001

Table 12: CI, Δ, and p for **online media (onl)** domain

Model	SO (CI)	SK		SKT		SKD	
		Δ	p	Δ	p	Δ	p
dschat	0.831 (0.813–0.846)	+12.65	< .001	+10.50	< .001	+8.55	< .001
gpt5m	0.880 (0.865–0.894)	+0.05	1.000	-1.60	.003	-0.40	.515
gemma2	0.919 (0.907–0.931)	+4.70	< .001	+3.80	< .001	+3.30	< .001
llama2	1.000 (0.998–1.000)	+0.00	< .001	+0.00	< .001	+0.00	< .001
llama3	0.951 (0.941–0.960)	-2.10	< .001	-1.95	< .001	-30.05	< .001
qwen2	0.785 (0.766–0.802)	-4.30	< .001	+0.50	.649	-9.10	< .001

Table 13: CI, Δ, and p for **legal (lgl)** domain

Domain	*gemma2_9b*	gemma3_12b	gemma3_27b	*llama3_8b*	llama3_70b	*qwen2_7b*	qwen2_72b
lit	79.0	70.0 (-9.0)	76.0 (-3.0)	82.5	78.0 (-4.5)	69.0	82.5 (+13.5)
jou	93.0	91.0 (-2.0)	95.0 (+2.0)	94.0	85.0 (-9.0)	61.5	91.0 (+29.5)
adv	65.5	77.5 (+12.0)	77.5 (+12.0)	73.5	69.5 (-4.0)	52.5	69.5 (+17.0)
acd	85.0	68.5 (-16.5)	91.5 (+6.5)	99.0	50.5 (-48.5)	30.5	79.5 (+49.0)
ins	91.5	91.0 (-0.5)	94.0 (+2.5)	90.5	84.5 (-6.0)	77.0	93.0 (+16.0)
com	79.5	71.5 (-8.0)	73.0 (-6.5)	81.0	73.0 (-8.0)	67.0	86.0 (+19.0)
onl	71.0	82.5 (+11.5)	77.5 (+6.5)	77.5	62.5 (-15.0)	61.5	79.0 (+17.5)
lgl	89.5	78.0 (-11.5)	74.5 (-15.0)	89.5	90.0 (+0.5)	71.5	90.5 (+19.0)

Table 14: Model Validation for **SO**

Domain	*gemma2_9b*	gemma3_12b	gemma3_27b	*llama3_8b*	llama3_70b	*qwen2_7b*	qwen2_72b
lit	94.5	88.5 (-6.0)	95.0 (+0.5)	82.5	90.5 (+8.0)	63.5	91.0 (+27.5)
jou	95.0	97.5 (+2.5)	99.5 (+4.5)	96.5	97.0 (+0.5)	70.0	96.0 (+26.0)
adv	87.0	89.5 (+2.5)	89.5 (+2.5)	84.5	85.5 (+1.0)	53.0	80.5 (+27.5)
acd	99.0	98.5 (-0.5)	99.0 (0.0)	99.5	99.0 (-0.5)	79.5	99.0 (+19.5)
ins	94.0	94.5 (+0.5)	97.0 (+3.0)	89.0	94.0 (+5.0)	82.0	95.5 (+13.5)
com	81.5	71.5 (-10.0)	77.0 (-4.5)	72.5	81.0 (+8.5)	55.5	87.5 (+32.0)
onl	62.0	85.5 (+23.5)	83.5 (+21.5)	58.0	75.5 (+17.5)	20.5	78.0 (+57.5)
lgl	95.5	95.0 (-0.5)	93.5 (-2.0)	89.5	96.5 (+7.0)	65.5	97.5 (+32.0)

Table 15: Model Validation for **SK**

Domain	*gemma2_9b*	gemma3_12b	gemma3_27b	*llama3_8b*	llama3_70b	*qwen2_7b*	qwen2_72b
lit	91.5	86.5 (-5.0)	94.0 (+2.5)	79.0	87.5 (+8.5)	74.0	92.5 (+18.5)
jou	93.5	96.5 (+3.0)	99.0 (+5.5)	95.5	95.5 (0.0)	79.5	94.5 (+15.0)
adv	94.5	87.0 (-7.5)	92.0 (-2.5)	69.5	85.0 (+15.5)	69.0	80.5 (+11.5)
acd	98.5	97.0 (-1.5)	99.0 (+0.5)	98.5	99.5 (+1.0)	91.0	99.5 (+8.5)
ins	92.5	92.5 (0.0)	97.0 (+4.5)	83.0	93.0 (+10.0)	81.0	94.0 (+13.0)
com	71.0	72.0 (+1.0)	70.0 (-1.0)	62.5	65.5 (+3.0)	41.5	86.5 (+45.0)
onl	75.0	88.5 (+13.5)	91.0 (+16.0)	66.5	69.5 (+3.0)	36.5	82.0 (+45.5)
lgl	94.5	93.5 (-1.0)	92.0 (-2.5)	89.5	95.5 (+6.0)	73.0	97.5 (+24.5)

Table 16: Model Validation for **SKT**

Domain	*gemma2_9b*	gemma3_12b	gemma3_27b	*llama3_8b*	llama3_70b	*qwen2_7b*	qwen2_72b
lit	90.5	77.5 (-13.0)	90.5 (0.0)	69.0	84.5 (+15.5)	55.5	89.0 (+33.5)
jou	89.5	98.0 (+8.5)	99.0 (+9.5)	88.5	93.0 (+4.5)	58.0	92.5 (+34.5)
adv	88.5	80.5 (-8.0)	86.5 (-2.0)	68.5	84.5 (+16.0)	24.5	72.0 (+47.5)
acd	98.5	95.5 (-3.0)	99.0 (+0.5)	96.0	99.0 (+3.0)	65.0	98.5 (+33.5)
ins	81.5	91.0 (+9.5)	96.0 (+14.5)	54.0	89.0 (+35.0)	57.5	89.5 (+32.0)
com	80.0	71.5 (-8.5)	76.0 (-4.0)	65.0	63.5 (-1.5)	40.0	83.0 (+43.0)
onl	59.0	84.5 (+25.5)	82.5 (+23.5)	59.0	33.5 (-25.5)	8.5	50.5 (+42.0)
lgl	94.0	94.0 (0.0)	91.5 (-2.5)	61.0	94.5 (+33.5)	62.0	96.0 (+34.0)

Table 17: Model Validation for **SKD**