



OPEN ACCESS

EDITED BY

Zhenguang Cai,
The Chinese University of Hong Kong, China

REVIEWED BY

Philippe Blache,
UMR7309 Laboratoire Parole et Langage (LPL),
France
Zhuang Qiu,
City University of Macau, Macao SAR, China

*CORRESPONDENCE

Pia Knoeferle
✉ pia.knoeferle@hu-berlin.de

RECEIVED 28 October 2025

REVISED 07 January 2026

ACCEPTED 19 January 2026

PUBLISHED 19 March 2026

CITATION

Knoeferle P (2026) ChatGPT-simulated sentence plausibility in event contexts, with teens, younger and older adults, in fiction and newspaper texts. *Front. Lang. Sci.* 5:1734306. doi: 10.3389/flang.2026.1734306

COPYRIGHT

© 2026 Knoeferle. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

ChatGPT-simulated sentence plausibility in event contexts, with teens, younger and older adults, in fiction and newspaper texts

Pia Knoeferle^{1,2,3*}

¹Faculty of Language, Literature, and Humanities, Humboldt-Universität zu Berlin, Berlin, Germany,

²Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, Germany, ³Einstein Center for Neurosciences Berlin, Charité - Universitätsmedizin, Berlin, Germany

The purpose of this study was to determine to what extent the large language model (LLM) would produce simulations that are close enough to human-based world knowledge to serve as pilot data for human experimentation: LLMs are developing rapidly, and if they become sufficiently accurate databases of human world knowledge, this would open up interesting opportunities for empirical research; with their advent we may have the opportunity of accessing a comprehensive model of world knowledge. This claim was assessed by simulating human plausibility ratings and their variation depending on (i) the presence vs. absence of an event description, (ii) the age of LLM-simulated participants (Pilot 1, Pilot 2, and Experiment 1a), and (iii) LLM-simulated participant expectations of distinct text sources/genres (Experiment 1b). In four pilot studies and two main experiments, ChatGPT-4o/5 plausibility ratings were simulated from the graphical user interface using written prompts, factorial designs, Latin-square counterbalanced lists, and $N = 200$ simulated participants per between-participant factor level. In this way, an experiment setup much like that for in-laboratory experiments with human participants was simulated. As a baseline, plausibility ratings generated via the LLM chat interface were compared against human plausibility ratings reported in prior research. Overall, ChatGPT produced simulated ratings that, on average, were higher for plausible than implausible sentences, and were higher when an event description supported the event conveyed by the target sentence. The model also revealed fine-grained differences depending on simulated participant age, context-sentence relations, and genre. These results can be used to guide the formulation of testable hypotheses for future research with human participants.

KEYWORDS

ChatGPT, inter- and intra-individual differences, large language models, plausibility ratings, simulations, world knowledge

1 Introduction

Much psycholinguistic research examined the role of world knowledge in language processing.¹ This examination assessed the time course with which such knowledge influences language processing via experiments, and conceptualized the mental representations of world knowledge in psycholinguistic theory development. For both

1 Other terms used for world knowledge in the literature are “encyclopedic” knowledge (Jackendoff, 1997, p. 31), “background” knowledge, and “experience” (Early, 1968, p. 2; Plaisance, 1928, p. 651f).

experimentally testing and conceptually modeling world knowledge, we must have an idea of what constitutes our world knowledge (e.g., who-does-what-to-whom, what instrument is used for an action, or where something is happening). One way in which this has been achieved is by means of plausibility rating studies in which participants' rated sentences such as (1) *The woman sailed the magazine...* and (2) *The woman edited the magazine...* for their plausibility (e.g., on a scale from 1 to 7; see, Pickering and Traxler, 1998, p. 944 and p. 959). Participants would rate sentences such as (1) as being much less plausible than sentences such as (2). From this type of outcome, we can conclude that the event expressed in (1) is less plausible and likely less often experienced than the event expressed in (2). Women editing magazines is thus part of our world knowledge, part of what we consider to be a plausible event; women sailing a magazine is a much more infrequent occurrence in our world (see, e.g., Connell and Keane, 2006, for a review and computational model of plausibility). This sort of knowledge can influence both text comprehension (e.g., Kintsch, 1988; Kintsch and Kintsch, 2005; McNamara, 2001; McNamara et al., 2004) and spoken language processing (e.g., Altmann and Kamide, 1999; Amsel et al., 2014; Chambers et al., 2002; Guerra et al., 2021; Hagoort et al., 2004; Kamide et al., 2003; Kutas and Federmeier, 2011; Nieuwland and van Berkum, 2006; Pescuma et al., 2023; Rommers et al., 2013; Sedivy et al., 1999; Tanenhaus et al., 1995).

Plausibility can further (a) be modulated by linguistic context, (b) vary with the age of the comprehender, and (c) vary with the text source. For instance, a discourse context can activate world (generalized event) knowledge and affect the processing of a subsequent target sentence (e.g., Metusalem et al., 2012). It can even change the interpretation of highly implausible sentences such as *The peanut was in love* toward a plausible meaning (e.g., if the peanut was described as having feelings in preceding discourse, Nieuwland and van Berkum, 2006, p. 1098). A linguistic context can even just be a single sentence: A preceding ("prime") sentence can, for instance, facilitate the processing of a following ("target") sentence (Arai et al., 2007; Knoeferle and Crocker, 2009; Pickering and Branigan, 1998). Such priming can be elicited via overlap in, and pre-activation of, noun and verb meaning or thematic role relations, thus drawing on event knowledge. Age can modulate the processing of sentence meaning and plausibility, too (Liu, 2024). In a sentence-judgment task older (61–78) and younger (18–26) adults rated sentences differing in semantic plausibility (plausible vs. implausible) and syntactic consistency (consistent vs. inconsistent) under high and low working memory load. Results reported by Liu (2024) revealed a stronger effect of semantic plausibility (and no reliable effect of syntactic consistency) in older adults for low working memory load. But when working memory load was high, semantic plausibility effects were stronger and syntactic consistency effects weaker in older adults compared to younger adults. These findings suggest that there may be an age-related focus on semantic plausibility effects and on reduced syntactic effects with increased working memory load. In addition, text source—often dubbed "genre"—may modulate plausibility ratings. In many text sources, such as narrations or newspaper texts, meaning is likely tied to the content of the text. But in some theories, such as Jakobson's genre-specific hypothesis (Jakobson, 1960, pp. 353 ff.), it is posited that

"Language must be investigated in all the variety of its functions. Before discussing the poetic function we must define its place among the other functions of language."

Further theories of literature, too, suggest that meaning may vary by genre (e.g., Culler, 1975; Hanauer, 1998). Against this background, Chen et al. (2016, p. 19) asked participants to read Chinese poems in which the final character was rhyming or semantically congruous vs. incongruous. Semantic incongruence elicited a classic N400 effect² only when the poem's rhyme scheme was preserved, but not when rhyme was violated. This indicates that in poetry reading, prosodic coherence (rhyme) guided whether semantic fit was evaluated. Such rhyme–semantic interactions suggest a genre-specific effect, where sound structure shapes meaning construction in ways not typically observed in prose. One interpretation of Chen et al.'s (2016) results is that

"what is considered to be meaning-based processing may depend in part upon the genre of a text."

The authors argue that this is

"important because a preponderance of the data we have in language processing comes from restricted linguistic genres." (p. 19)

In summary, human world knowledge plays an important role in language-related processes and such effects can—at least in some cases—be modulated by event contexts, participant age, and text genre.

Recently, large language models (LLMs) such as ChatGPT have been examined as models of world knowledge. Kauf et al. (2023) reported that LLMs were good at attributing higher likelihood to possible over impossible events such as a laptop buying a teacher. But the LLMs were less apt at attributing higher likelihood to likely (a nanny tutoring a boy) over unlikely events (a boy tutoring a nanny; Table 1 in Kauf et al., 2023). This difference seems to highlight a gap between how humans represent and evaluate possible/impossible and likely/unlikely events and how LLMs represent these plausibility differences (see also Demszky et al., 2023, on concerns regarding the use of LLMs for psychology research; see Cai et al., 2024, on differences and similarities between humans and ChatGPT in language processing). But LLMs are developing rapidly, and if they were sufficiently accurate databases of human world knowledge, this would open up interesting opportunities for empirical research:

"A few years ago, it would have been a challenge to capture world knowledge within a culture and for a population such as young healthy adults. But with the advent of large

² The "N400 effect" is a difference in averaged brain waves with more negative-going mean amplitude event-related brain potentials for semantically incongruous than congruous stimuli (see Kutas and Federmeier, 2011, for a review).

language models we have the opportunity of accessing a very comprehensive model of world knowledge.” (Knoeferle, 2025, p. 26)

This would be particularly useful if one could conduct language experiments examining world knowledge using the graphical user interface.

1.1 The present research

The present research builds on these findings and asks to what extent one LLM— ChatGPT (versions 4o and 5)— captures world knowledge reflected in simulated plausibility ratings and its variation depending on (i) the presence vs. absence of an event description, (ii) the age of LLM-simulated participants (Pilot 1, Pilot 2, and Experiment 1a), and (iii) LLM-simulated participant expectations of distinct text sources (Experiment 1b). This work will therefore contribute to an agenda that fills “gaps in understanding how computational advances correspond to actual cognitive mechanisms, how language processing varies across individuals and groups, and the extent to which empirical paradigms effectively address theoretical questions” (see research topic description, <https://www.frontiersin.org/research-topics/63095/insights-in-psycholinguistics-2025>).

In using ChatGPT-produced plausibility ratings, it seems to be good practice to have an understanding of the LLM’s definition of plausibility ratings when it is explicitly probed. Accordingly, in conducting pilot studies, I asked ChatGPT-4o to provide a definition of plausibility (*How do you define “plausibility”?*). ChatGPT responded that plausibility refers to how believable or realistic a sentence is, based on how naturally it aligns with logic, grammar, and real-world knowledge. In the context of sentence stimuli, it defined plausibility based on three factors:

1. Grammatical coherence—Does the sentence follow standard syntactic rules? If it has misplaced modifiers, awkward phrasing, or ambiguous structures, it may seem less plausible.
2. Semantic clarity—Does the sentence make sense in terms of meaning? A sentence that describes an impossible or highly unlikely event (e.g., *The moon drank coffee at midnight.*) would be implausible.
3. Pragmatic realism—Does the sentence describe something that could reasonably happen in the world? If an action is unusual but still possible, it may receive a lower plausibility rating, whereas common, expected scenarios score higher.

By comparison with the definition by ChatGPT, extant research involving humans teased apart plausibility from grammaticality/syntax (Lau et al., 2017, p. 1204; Pickering and Traxler, 1998; Thornton and MacDonald, 1997). Others have not made such differentiation: plausibility

“in its most general form, can be defined as the acceptability or likelihood of a situation or a sentence describing it, as a whole.” (Matsuki et al., 2011, p. 926)

TABLE 1 Average plausibility ratings (Mean \pm SD) across baseline files.

Sentence	Mean	SD
A pork roast is a desert.	1.0	0.0
At many nightclubs one can listen to techno music.	6.0	0.0
At the North Pole one can find giraffes.	1.0	0.0
At the zoo, lions can look at children in cages.	1.5	0.7
Aunts often may like to give gifts to their nieces.	6.0	0.0
Babies feed their moms.	2.0	0.0
Ballerinas like to dance.	7.0	0.0
Cake contains flour and water and sugar and eggs.	7.0	0.0
Cars can drive fast on the motorway.	7.0	0.0
Clouds are always pink.	1.0	0.0
In the courtroom one must be respectful.	7.0	0.0
In a soccer match, 5 people play with a toilet paper roll.	2.0	0.0
Meaning full boys are unhappily going.	2.0	0.0
New Year’s Eve is celebrated in June.	1.0	0.0
On a playground snails are using the slide.	1.0	0.0
The police officers should follow the law.	7.0	0.0
The sand is whitish.	6.0	0.0
The student gave the teacher a bad grade in the English exam.	3.0	0.0
The sun is shining and it is dry.	7.0	0.0
Today is lovely weather and the birds are chirping.	6.0	0.0

It has further been defined as “the degree of fit between a given scenario and prior knowledge” (Connell and Keane, 2006, p. 98); see also Wertgen and Richter (2020).

That said, baseline sentences that ChatGPT-4o and 5 were asked to rate for plausibility on a scale from 1 (very implausible) to 7 (very plausible) elicited ratings that one might expect to see from human participants (see Table 1 for averages of plausibility ratings that ChatGPT provided across nine runs of the listed sentences; the very low standard deviation values suggest high consistency across these runs).

It is worthwhile to assess to what extent simulated plausibility ratings reflect the definition given by ChatGPT and to what extent they reflect a distinction of grammatical coherence (in the structure of sentences) and plausibility (referring to whether the sentence describes something that would happen in the world). To that end, the pilot experiments (Pilots 1a, 1b, 2a, 2b) contained a within-experiment sentence structure factor (canonical vs. non-canonical) to assess the interactions of plausibility ratings with sentence structure. This provides insight into the extent to which plausibility ratings are modulated by difficult-to-understand and infrequent sentence structure and thus can be used to dissociate the effects of sentence structure from plausibility.

The main experiments examined the effects of context presence and age (Experiment 1a), and, using identical target sentences and event contexts as in Experiment 1a, of genre (Experiment 1b) on simulated plausibility ratings. Experiments 1a and 1b used a new set of materials compared to the pilot studies and instead of a within-participants sentence structure manipulation (e.g., object vs. subject-first), the design used German object-initial sentences only. Experiments 1a and 1b contrasted plausibility ratings for these in different sentence-context relations within participants (see Section 3, Table 4). This was done to assess whether LLM-simulated plausibility ratings for relatively low-plausibility events in non-canonical object-initial sentences in German could be boosted by means of an event context description; and whether it would vary depending on how the context—if present—relates to the sentence, what participant age the LLM assumes for the simulated ratings, or what genre the text is presumed to originate from.

In the following, first a summary is given of the hypotheses of the reported research for the between-participant variables: (i) presence vs. absence of an event description, (ii) the age of LLM-simulated participants (Pilot 1, Pilot 2, and Experiment 1a), and (iii) LLM-simulated participant expectations of distinct text sources (Experiment 1b). Subsequently, I describe the materials, design (including the two-level within-participants sentence-structure manipulation), and methods for the pilot studies and report their results. This is followed by a description of Experiments 1a and 1b. To give a baseline for the LLM-based simulations, I also compare the model-generated ratings with existing human data from Kauf et al. (2023).

1.2 Hypotheses

1.2.1 Event context

To illustrate the hypothesis that event descriptions could influence plausibility ratings, consider that a discourse context can activate world (generalized event) knowledge and affect the processing of a subsequent target sentence (Metusalem et al., 2012). It can even change the interpretation of highly implausible sentences such as *The peanut was in love* toward a plausible meaning (e.g., if the peanut was described as having feelings in preceding discourse, Nieuwland and van Berkum, 2006, p. 1098). A preceding (“prime”) sentence influenced the processing of a following (“target”) sentence (Arai et al., 2007; Pickering and Branigan, 1998; Knoeferle and Crocker, 2009), likely via overlap in noun and verb meaning, structural, and/or thematic role relations. If an LLM also considers linguistic context (in the form of concise event contexts) as a modulator of the plausibility rating for a target sentence, then we should see higher plausibility ratings for implausible target sentences enriched with (vs. without) event context.

1.2.2 Participant age

Another factor modulating plausibility ratings could be participants’ age. When asking ChatGPT-4o about differences regarding language and world knowledge across the lifespan, its responses were individuals aged 16 to 18 are in a transitional

phase and balance casual, social-driven language is combined with growing professional proficiency; and that world knowledge is shaped by education, internet, and increasing social awareness. For those aged 18–31 traditional knowledge is individuals combined with modern, digital, and pop culture; communicate efficiently, stay updated, and navigate global issues with digital fluency. People aged between 55 and 65 tend to use more formal and structured language, and rely less on digital slang and internet trends; their world knowledge is shaped by lived experiences, traditional media, and professional expertise, giving them a deep but sometimes cautious perspective on modern developments.

If age has no effect, ratings should be the same for all three age groups. Alternatively, age may interact with context and/or sentence structure, perhaps reflecting differences in the simulated world and language knowledge of the participants. Maybe the life experience of 55–65 year olds—as encoded in ChatGPT’s knowledge base—renders them more likely to rate even implausible events as more plausible; if so, then simulated ratings for the older group should be higher than for the younger groups. Alternatively, young adults are in a more creative space and this sort of greater creativity and flexibility of young (vs. older) minds (that may be more “set”) might be encoded in ChatGPT’s knowledge base. To the extent that this reasoning holds, we should see higher plausibility ratings for low-plausibility sentences in the younger than the older ChatGPT-simulated participant group. Based on Liu (2024) finding that semantic plausibility effects were stronger in older than younger adults, to the extent that their findings generalize, semantic plausibility differences (with context present vs. no context, on the assumption that context boosts plausibility) should be larger in older than younger adults. In addition, one might expect no changes of plausibility ratings depending on sentence structure for the older participants given the reported weaker syntactic consistency effects for this group in Liu (2024).

1.2.3 Text source

Depending on the source of a text, different language use is customary, and this is part of our linguistic and world knowledge (Biber and Conrad, 2009; Goulart et al., 2020, p. 436, on register as non-conventional functional language use compared with genre as containing more conventionalized language). Consider differentiating fictive from non-fictive texts: For instance, in Goethe’s poems, object-initial word order may not be unexpected (“Und nichts zu suchen Das war mein Sinn,” literal translation: “And nothing to search This was my intention,” Atkins and Kastner, 1902, p. 17f.). In Charlotte Brontë’s poems, unusual word choices may be expected, and even non-canonical word order, such as “The yoke put on, the long task done,—I am, as it is bliss to be, ...” (retrieved from <https://interestingliterature.com/2019/11/the-best-charlotte-bronte-poems-everyone-should-read/>, December 8, 2025). In a non-fiction text such as a newspaper, the same kinds of unusual word order and word choices would arguably not be expected to the same extent. Some results suggest that genre expectations influence language comprehension and what one retains from a text (e.g., Zwaan, 1994). For instance, participants reading on the assumption of processing literary texts had longer reading times and better

memory for surface information (discriminating between verbatim and paraphrased text sentences, Experiment 1 in Zwaan, 1994); their memory for situational information (discriminating plausible from implausible situation inferences) was worse than readers expecting to be reading text from news stories. To the extent that this is also reflected in ChatGPT's world knowledge and plausibility ratings, its plausibility ratings should differentiate more clearly between plausible and implausible sentences for non-fictional (newspaper) than fictional (literary) text sources (see also Chen et al., 2016, for evidence on genre effects for silent reading).

2 Pilot studies

English was chosen for the initial piloting; subsequent piloting ensured that the prompting method generalized to German stimuli since German was the language for the main experiments. Pilots 1a and 1b used English as the experiment language. For Pilots 2a and 2b, the experiment language was German. Pilots 1a, 2a included an event context with the target sentence, while Pilots 1b, 2b did not.

2.1 Materials

For Pilot 1, the sentences were taken from Experiment 1 in Knoeferle and Crocker (2006). The event contexts were generated by the author from the images in Knoeferle and Crocker (2006) and rendered the event in short form with canonical word order. For Pilot 2, the sentences were taken from Knoeferle et al. (2005) and the event contexts were also generated by the author from the corresponding images. A copy of the materials (event contexts and target sentence stimuli labeled myitems_PEng.csv and myitems_PGer.csv) is available on the Open Science Framework (OSF) repository at <https://osf.io/u2tqr/overview>. The stimuli are also listed in Knoeferle (2005), Appendix A.3, <https://publikationen.sulb.uni-saarland.de/handle/20.500.11880/23525>.

The materials for Pilot 1 (English) contained a main-clause/reduced relative clause ambiguity illustrated by in Examples A (1.a), (1.b). The materials for Pilot 2 (German) contained a local structural (case marking) ambiguity illustrated in Examples B (1.a), (1.b). In addition to the two sentence versions a) and b) given for Examples A and B, each experimental item contained a further two sentences to counterbalance who-does-what-to-whom and thus the thematic fit of individual agents/patients with individual verbs. For instance, to counterbalance Examples 1.(a) and 1.(b), the experimental item also contained Examples 2.(a) *The ballerina sketched apparently the fencer in the white suit.* and 2.(b) *The ballerina splashed apparently by the cellist sketched the fencer.* The full experiment including counterbalancing, thus included four sentences and two image descriptions. The counterbalancing ensures that a better thematic fit of, say, ballerina as sketching-agent or fencer as splashing-agent (than vice versa) would not bias the ratings (both agents occur once as agent/patient of splash/sketch in the main and reduced-relative clause).

The same reasoning holds for the German stimuli. In addition to the sentences in Examples B (1.a), (1.b), each item contained

two further sentences for counterbalancing reasons. For Examples B (1.a), (1.b), these were 2.a) *Die Prinzessin malt offensichtlich den Fencer* ["The princess-amb./subj. paints apparently the fencer (obj.)"] and 2.b) *Die Prinzessin wäscht offensichtlich der Pirat* ["The princess-amb./obj. washes apparently the pirate (subj.)"]. Thus, in a Latin-square design, ratings are obtained for princess-painting-fencer, princess-washing-fencer, fencer-painting-princess, and fencer-washing-princess. Thematic role relations are thus fully counterbalanced within each item across lists.

Contextual manipulation: In the context-based versions of the pilots, each of the target sentences was paired with a contextual description. For instance, the English sentences given in Examples A (1.a), (1.b) were paired with *ballerina splashing cellist, fencer sketching ballerina*. The German sentences in Examples B (1.a), (1.b) were each paired with *Prinzessin wäscht Pirat, Fencer malt Prinzessin* (princess-washes-pirate, fencer-paints-princess). This means that, in this context, the role relations of the target sentence were restated in canonical (agent-patient) order. Understanding, for instance, *The ballerina sketched apparently by the fencer splashed the cellist* could be facilitated when the context contains *fencer-sketching-ballerina or ballerina-splashing-cellist*. At the same time, resolving the ordering differences could reduce the plausibility rating compared with when both target and contextual descriptions contain the same ordering of agent-patient. In summary, for the target-context relations, the same order (for the canonical sentences) or a different order of role fillers (for the non-canonical/low-frequency structures), or differences in the amount of semantic overlap, could influence plausibility ratings. Low plausibility ratings could then be due to mismatches between contextual description and the target sentence.

A Example

- 1.a) The ballerina splashed apparently the cellist in the white shirt. (mainclause)
- 1.b) The ballerina sketched apparently by the fencer splashed the cellist. (reducedrelativeclause)

B Example

- 1.a) Die Prinzessin wäscht offensichtlich den Pirat. (subject—verb—object)
literal translation: "The princess (amb.-subj.)-washes-apparently-the pirate (object)."
- 1.b) Die Prinzessin malt offensichtlich der Fencer. (object—verb—subject)
literal translation: "The princess (amb.-obj.)-paints-apparently-the fencer (subject)."

Post-hoc, the (i) semantic and surface structural aspects of the target sentences and (ii) the semantic relation between the context and the target sentence were qualified. To this end, four complementary measures were computed for each context-target sentence pair to characterize surface-level structure and semantic expectedness. Sentence compression ratio indexed structural variability by dividing the byte length of the sentence after gzip compression by its raw character length, with higher values indicating greater structural complexity or lower compressibility. Sentence-description cosine similarity quantified semantic expectedness as the cosine similarity between TF-IDF-weighted

vector representations of the sentence and its corresponding context computed within each dataset; higher values indicate greater overlap in informative lexical content and closer semantic alignment of the target sentence with the event context. To capture relative expectedness among different sentences, within-event context z -scored similarity was computed by z -scoring sentence-event description cosine values, producing a standardized index of how much a sentence exceeded or fell below the mean semantic alignment for that event context. Finally, within-event context expectedness rank-ordered sentences for each image by their cosine similarity (with rank 1 indicating the highest similarity), providing a nonparametric measure of relative expectedness. Together, these measures permit expectedness effects to be assessed both in absolute terms (raw cosine similarity) and relative terms (within-event context standardization and rank), while controlling for surface-level structural variation. The analysis was conducted twice, each time giving ChatGPT 5.2 the stimuli .csv files as input and asking it to compute the four dependent measures. The files for this analysis can be found in the OSF repository for the article, folder “Stimuli_analysis.”

Across the original and replication runs, sentence compression ratio showed no reliable effects of condition in any dataset, confirming that the materials were well matched for surface-level structural complexity. In the English pilot study materials, sentence-event context cosine similarity robustly distinguished conditions in both rounds of analysis, with reduced relative clauses showing substantially higher semantic similarity to the image description than main clauses [replication: $M(\text{red. rel.}) = 0.4$ vs. $M(\text{mainclause}) = 0.3$; $t = -4.18$, $p < 0.001$; permutation $p < 0.0001$; mixed-effects $\chi^2(1) = 26.19$, $p < 0.001$]. In the German pilot study materials, cosine similarity showed a reliable condition effect in the original run ($ps < 0.05$) but not in the replication ($p > 0.5$). Within-context z -scored similarity and rank measures were degenerate and therefore treated descriptively in the two-condition pilot datasets. What these measures capture is that more context-target sentence overlap (reduced-relative clause vs. main clause sentences in the English pilot) resulted in higher cosine similarity ratings. In contrast, for the German pilot, the object- and subject-initial sentences had comparable lexical-semantic overlap with the event context. [Figure 1](#) plots the results for the replication runs.

2.2 Event context and age group effects

If only the long-term knowledge in ChatGPT determines plausibility ratings, then describing an event as an immediate context shouldn't affect them. If so, then paired sentences (event description and target sentence) should not differ in plausibility ratings from target sentences without the event description. Alternatively, if describing an immediate event influences the assessment of paired language as more plausible than without the description, then ChatGPT sentence plausibility ratings should be higher when paired with the event description (Pilots 1a/2a) than on their own (Pilots 1b/2b).

In addition, ChatGPT was asked to simulate the plausibility ratings of human participants in different age groups (16–18 year olds vs. 18–31 vs. 55–65 year olds). This is to see if plausibility

ratings differ for these age groups, linking to prior results and generating testable predictions for experiments with actual human speakers. To the extent that the life experience of 55–65 year olds—as encoded in ChatGPT's knowledge base—renders them more likely to rate even implausible events as more plausible, simulated ratings for the older group should be higher than for the younger groups. Alternatively, young adults are in a more creative space and this sort of greater creativity and flexibility of young (vs. older) minds (that may be more “set”) might be encoded in ChatGPT's knowledge base. To the extent that this reasoning holds, we should see higher plausibility ratings for low-plausibility sentences in the younger compared to than in the older ChatGPT-simulated participants. Recall further, [Liu et al. \(2024\)](#)'s finding that semantic plausibility effects were stronger in older adults than in younger adults; to the extent that this generalizes, semantic plausibility differences (with context present vs. no context, on the assumption that context boosts plausibility) should be larger in older than in younger adults. In addition, one might expect no changes of plausibility ratings depending on sentence structure for the older participants given the reported weaker syntactic consistency effects for this group in [Liu et al. \(2024\)](#).

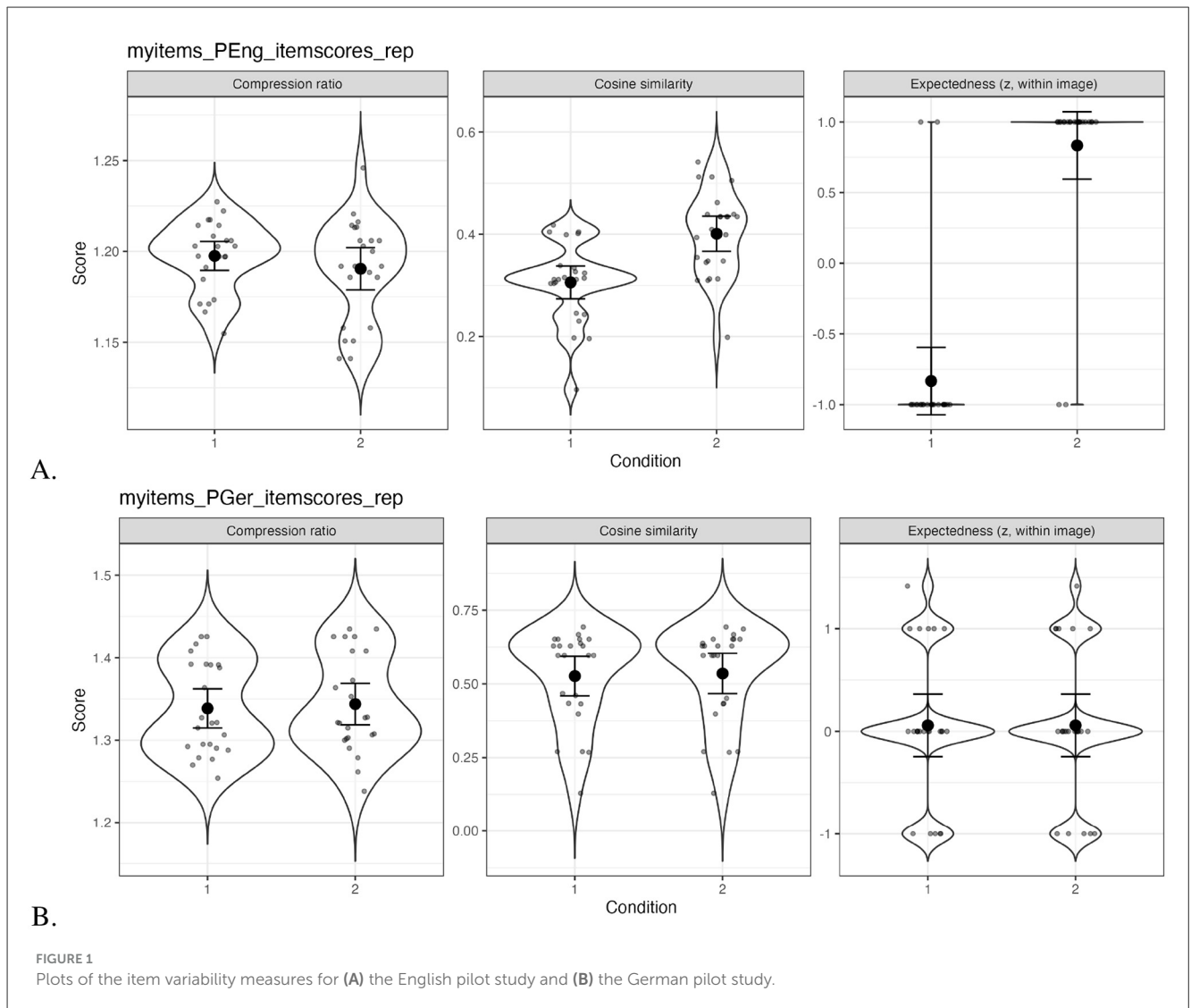
2.3 Sentence structure effects

The target sentence event relations in the pilot studies were non-stereotypical and had low-plausibility. If ChatGPT captures this, then the sentences presented alone should have low plausibility ratings (e.g., 1–3.5 on a Likert-like scale from 1 to 7; 1 indicating “not plausible at all” to 7 indicating “highly plausible”). Differences in sentence structure between conditions a and b could affect plausibility assessments (see Examples A and B). **Pilot 1:** In English, condition b sentences had a more complex reduced relative clause structure compared to condition a sentences with a main clause (Example 1). Condition b sentences had temporary ambiguity, where the second verb could be a simple past or a past participle, making them temporarily interpretable as the active verb in a main clause fragment or as the past participle in a reduced relative clause. **Pilot 2:** In German (Example 2), condition b sentences had a less frequent and non-canonical object-verb-subject structure compared to condition a sentences with a subject-verb-object structure. If this difference in sentence structure and associated processing ease/difficulty modulates simulated plausibility assessments, we might see lower plausibility ratings for sentences in condition b than a (directed hypothesis).

2.4 Methods and design

2.4.1 Design

The design was a mixed design with two between-participant factors: age (16–18, 18–31, 55–65 years olds) and context (context vs. no context) and one within-participants factor (sentence) condition (two levels: main-clause vs. reduced-relative clause in Pilots 1 and subject-verb-object vs. object-verb-subject in Pilots 2). There are two sentence versions per item (each with condition a and b, recoded as a = 1 and b = 2 in the prompts). These



versions were created in the original experiments to counterbalance thematic role relations. For instance, a ballerina splashing a cellist was part of the main-clause condition and a ballerina being splashed by the cellist was part of the reduced-relative clause condition. The ballerina and the cellist were thus part of a splashing event in both thematic roles (agent; patient). As a result, plausibility rating differences cannot be due to differences in the lexical content or agent/patient-verb fit of the sentences.

2.4.2 Prompt development

Care was taken to ensure that the prompts for the main experiments could be executed by ChatGPT as intended. For Pilot 1a, 19 runs of the prompts were conducted until the instructions yielded the desired results. Adjustments across runs included asking ChatGPT to wait for a “Go” signal (later simplified), until all stimuli had been uploaded; double-checking that image description versions were allocated as intended to sentence versions; logging out in-between runs, clearing browser history and ChatGPT session memory via settings to avoid the history influencing ratings on a new run in the ChatGPT-4o version; this was no longer necessary

for ChatGPT5 for which this was controllable via the settings; scaling the pilot from $N = 1$ rating provided by ChatGPT to $N = 200$ simulated participant ratings; verifying the normality of the rating distributions; creating a Latin-square item-condition allocation file; and asking ChatGPT for increasingly more detailed outputs in separate files. For Pilot 1b, 10 runs were conducted. For Pilot 2a, five runs were conducted and for Pilot 2b, one run. All runs except the final run were conducted using ChatGPT-4o (2024/2025). The final run for each pilot was conducted in September 2025 using ChatGPT5. The detailed prompts are available at <https://osf.io/u2tqr/overview>.

Per-sentence plausibility ratings and rating averages by condition were obtained, these once for ChatGPT simulating just one participant, and in a second version from ChatGPT with the request that the LLM simulate 200 participants; $N = 200$ is a sensible number of participants for psycholinguistic plausibility ratings in an experiment with two sentence conditions and planned between-experiment comparisons (see Brysbaert, 2019).

Post-data generation, it was checked using ChatGPT5 that the data were indeed from independent participants (regarding their variability and potential duplicates). This held for all datasets with

only few duplicates (≤ 20) across all tested datasets (6×200 participants for each pilot study; 1,200 participants for Experiment 1a, and 400 participants for Experiment 1b).

2.4.3 Analysis

Here, the analysis of Pilots 1 and 2 is reported. The focus will be on the between-participant factors age (16–18, 18–31, and 55–65 years of age), context (context vs. no context), and condition (two sentence structures that differ in processing difficulty, e.g., subject–verb–object compared with object–verb–subject).

The pilot studies were conducted and their results analyzed before pre-registration (<https://doi.org/10.17605/OSF.IO/GEJRP>). But after pre-registration it became clear that the generated data contained little participant variability beyond the simulated age variability. For that reason, the pre-registration data files were again inputted to ChatGPT and ChatGPT was to regenerate the plausibility ratings with more participant variability using the prompt:

“Here is a file, pilot._all_between.csv: The column PlausibilityRating contains plausibility ratings from participants. Context_label and age_label tell you what context the participants experienced and what their age is respectively; ParticipantID tracks participant number, ItemID tracks item number. The ratings thus come from participants in different age groups. But that is the only variability in the ratings. Please re-generate the ratings, bearing in mind the age of the participant but also add variability that simulates other likely participant characteristics (always bearing in mind the age). Re-output that file as pilot._all_between_par.csv.”

The results pattern is highly similar to the reported regenerated Pilot ratings. An updated .R script was used in the analysis reported below. The updated script included model comparison starting from the most complex model justifiable by the design (Barr et al., 2013); these files are available on OSF as Pilot1rev and Pilot2rev (<https://osf.io/u2tqr/overview>).

Data were analyzed using linear mixed-effects regression models (lme4, Bates et al., 2015) with plausibility ratings as the dependent variable. The primary fixed-effects structure included condition, Age group, and Context, along with their interactions, with factors coded using sum contrasts. Random-effects structures were specified to allow for by-participant and by-item intercepts and, where supported, random slopes for condition. Three candidate models were fit (maximal, reduced with slopes by participants only, and intercepts-only), with model selection based on maximum-likelihood Akaike Information Criterion (AIC) comparisons. The final model was re-fit using restricted maximum likelihood (REML) for inference, and singularity checks were conducted to assess variance components. Type III ANOVAs (Kuznetsova et al., 2017) provided significance tests for fixed effects, and estimated marginal means with 95% confidence intervals were visualized for the condition \times age \times context interaction. Model summaries, random-effects variance components, and model-selection statistics were exported in both tabular (LaTeX, CSV) and graphical formats to ensure reproducibility and transparent reporting. Given the ordinal nature of the data, for the main

experiments, in addition an ordinal cumulative link mixed model with random intercepts for participant and item was fitted to verify that treating plausibility ratings as ordinal rather than continuous did not change the fixed-effect conclusions. Results were qualitatively similar when re-analyzed using cumulative link mixed models. In addition, the lmer was simplified to eliminate singular fit; comparing results for the more complex and reduced lmer showed fixed effects were stable.

Pilot 1. The best-fitting model included random intercepts for participants and items only. Both variance components were non zero, and Singular() returned FALSE, indicating that this model was not singular.

The selected model for Pilot 1 (English) was: $PlausibilityRating \sim Condition + age_label + context_label + (1|ParticipantID) + (1|Item) + Condition:age_label + Condition:context_label + age_label:context_label + Condition:age_label:context_label$.

Pilot 2. A linear mixed-effects model with random intercepts and by-condition slopes for both participants and items was fitted. Inspection of the variance-covariance estimates revealed that the variance of the item intercept was estimated at zero, while the participant intercepts and slopes, as well as item slopes, showed nonzero variance. According to the isSingular() diagnostic in lme4, this pattern indicates that the model is formally singular. However, because the slope variances were non zero and the model converged stably, this structure was retained following recommendations to preserve the theoretically justified random-effects specification (Barr et al., 2013; Bates et al., 2015). The selected model for Pilot 2 (German) was: $PlausibilityRating \sim Condition + age_label + context_label + (1 + Condition|ParticipantID) + ((1|Item) + (0 + Condition|Item)) + Condition:age_label + Condition:context_label + age_label:context_label + Condition:age_label:context_label$.

2.5 Pilot study results

2.5.1 Descriptive results

Figure 2 illustrates the main effects and interactions. For the English data (Pilot 1), Figure 2A. shows low to mid plausibility ratings (≤ 4 on a scale from 1 to 7 with 7 being the most plausible). With context, ratings had an average of $M = 3.26$ compared with $M = 3.13$ with no context. The highest ratings came from the 18–31 year-olds; in contrast, 16–18 and 55–65 year olds had lower ratings. When context was present, it did not affect ratings for the two sentence conditions in the 18–31 year olds; 16–18 year olds, however, rated main clauses (Condition 1) higher than reduced relative clauses (Condition 2) with context; the older adults in contrast, rated reduced relative clauses (Condition 2) higher than main clauses (Condition 1) with context. When no context was given, all age groups tended to give higher ratings to the more frequent main clause structure; this tendency was strongest in the 55–65 year olds.

For the German data (Pilot 2), Figure 2B. illustrates that ratings were higher with context ($M = 4.64$), than with no context ($M = 3.84$). The highest ratings came from the older adults, and lower

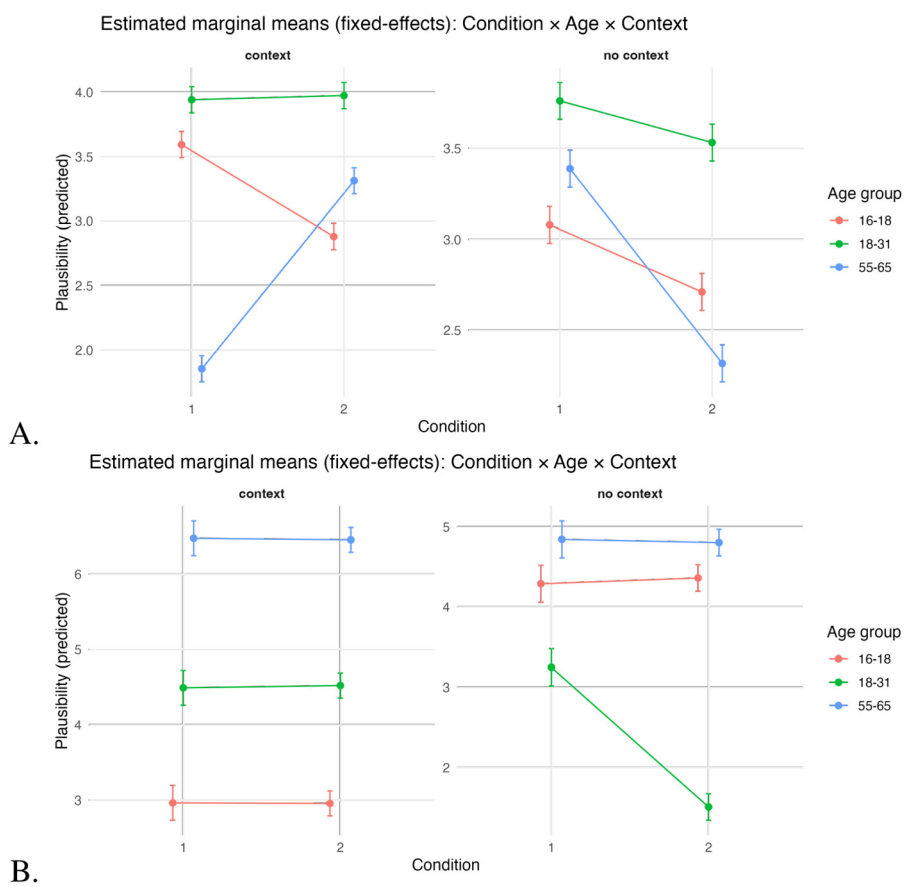


FIGURE 2 (A) Pilot 1—English; (B) Pilot 2— German. Condition 1 is the main clause (Pilot 1)/subject–verb–object condition (Pilot 2), and Condition 2 is the reduced relative clause (Pilot 1)/object–verb–subject condition (Pilot 2). The plot illustrates the significant three-way interaction: age-related differences in plausibility judgments varied across conditions, and the magnitude of these differences depended on whether supportive context was available. Estimated marginal means ($\pm 95\%$ confidence intervals) are plotted as a function of condition, age group, and context. Points and lines show fixed-effect predictions from the selected mixed-effects model; error bars indicate 95% confidence intervals derived from the fixed-effects covariance matrix. Colors distinguish age groups, and facets separate sentences presented with vs. without contextual support.

ratings were given by the 18–31 year olds and the lowest by the 16–18 year olds. When context was present, it did not seem to affect ratings of subject–verb–object compared to object–verb–subject sentences in the 18–31 year olds. In contrast, without context, the 18–31 year olds gave much lower ratings for object-initial sentences than subject-initial sentences, for which their ratings were approximately the same as when context was present. This tendency and differentiation by context was not seen in the two other age groups.

2.5.2 Inferential analysis: Pilot 1 (English)

For Pilot 1 (English), a linear mixed-effects model was fitted to plausibility ratings with condition (within-participants), age group (between-participants), and context (between-participants) as fixed effects, and random intercepts for participants and items. Model comparison indicated that this intercept-only structure provided the best fit (AIC = 48,547.0), with more complex random-slope models performing slightly worse ($\Delta AIC = 4-7$). The Type III ANOVA (see Table 2) revealed significant main effects of condition, $F_{(1,13,183)} = 49.10, p < 0.001$; age, $F_{(2,1,194)} = 467.13, p < 0.001$; and

context, $F_{(1,1,194)} = 18.90, p < 0.001$. These were further qualified by significant interactions of condition \times age, $F_{(2,13,183)} = 102.22, p < 0.001$; condition \times context, $F_{(1,13,183)} = 372.45, p < 0.001$; and age \times context, $F_{(2,1,194)} = 45.20, p < 0.001$. There was a robust three-way interaction of condition \times age \times context, $F_{(2,13,183)} = 428.21, p < 0.001$. The effect of condition was positive overall, but it increased with age, and this age-related change was itself modulated by context. Thus, plausibility ratings were shaped not only by condition but by the combined influence of age and context, yielding a reliable three-way interaction.

2.5.3 Inferential analysis: Pilot 2 (German)

For Pilot 2, a linear mixed-effects model was fitted to plausibility ratings with condition (within-participants), age group (between-participants), and context (between-participants) as fixed effects, and random intercepts and slopes for condition by participants and items. Model comparison indicated that the full random structure provided the best fit (AIC = 46,944.9), substantially outperforming simpler models ($\Delta AIC \geq 34.8$). The Type III ANOVA (see Table 3) revealed significant main effects

TABLE 2 Pilot 1—English: Type III ANOVA results from the selected linear mixed-effects model predicting plausibility ratings.

Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(> F) ^a	Effect
78.914	78.914	1	13,183	49.099	0	Condition
1,501.573	750.787	2	1,194	467.132	0	Age_label
30.384	30.384	1	1,194	18.904	0	Context_label
328.585	164.293	2	13,183	102.221	0	Condition:age_label
598.618	598.618	1	13,183	372.454	0	Condition:context_label
145.285	72.643	2	1,194	45.198	0	Age_label:context_label
1,376.463	688.232	2	13,183	428.211	0	Condition:age_label:context_label

The model included fixed effects of condition, age group, and context, as well as all two-way and three-way interactions, with random intercepts for participants and items. Reported statistics correspond to *F*-tests with Satterthwaite approximations for denominator degrees of freedom. Significant effects indicate that plausibility ratings varied systematically across conditions, and that both age and context moderated the size of the condition effect, culminating in a significant three-way interaction.

^aProbability (Pr) of observing a value greater than the obtained *F*-statistic under the null hypothesis.

TABLE 3 Pilot 2—German: Type III ANOVA results from the selected linear mixed-effects model predicting plausibility ratings.

Sum Sq	Mean Sq	NumDF	DenDF	F-value	Pr(> F) ^a	Effect
45.509	45.509	1	10,588	31.987	0	Condition
6,703.204	3,351.602	2	1,193.879	2,355.741	0	Age_label
1,091.740	1,091.740	1	1,193.879	767.352	0	Context_label
583.133	291.566	2	9,857.977	204.933	0	Condition:age_label
289.966	289.966	1	9,857.977	203.808	0	Condition:context_label
4,053.137	2,026.569	2	1,193.879	1,424.414	0	Age_label:context_label
643.074	321.537	2	9,857.977	225.999	0	Condition:age_label:context_label

The model included fixed effects of condition, age group, and context, as well as all two-way and three-way interactions, with random intercepts for participants and items. Reported statistics correspond to *F*-tests with Satterthwaite approximations for denominator degrees of freedom. Significant effects indicate that plausibility ratings varied systematically across conditions, and that both age and context moderated the size of the condition effect, culminating in a significant three-way interaction.

^aProbability (Pr) of observing a value greater than the obtained *F*-statistic under the null hypothesis.

of condition, $F_{(1,10.59)} = 31.99$, $p < 0.001$; age, $F_{(2,1,193.88)} = 2,355.74$, $p < 0.001$; and context, $F_{(1,1,193.88)} = 767.35$, $p < 0.001$. These effects were further qualified by significant two-way interactions of condition \times age, $F_{(2,9,857.98)} = 204.93$, $p < 0.001$; condition \times context, $F_{(1,9,857.98)} = 203.81$, $p < 0.001$; and age \times context, $F_{(2,1,193.879)} = 1,424.414$, $p < 0.001$. Crucially, there was also a significant three-way interaction of condition \times age \times context, $F_{(2,9,857.977)} = 225.999$, $p < 0.001$. Inspection of the fixed-effect estimates indicated that the effect of condition varied by age group, with older participants showing a reduced condition effect relative to younger participants. Moreover, this age-related difference was modulated by context, such that the size and direction of the condition effect differed across contexts.

2.6 Discussion: LLM plausibility rating simulations

In terms of evaluating the LLM-generated plausibility ratings, recall the original hypotheses: Plausibility ratings should all be low without context for the pilot studies (e.g., below 3.5). If context could boost plausibility ratings, then these should be higher for implausible sentences with context than without. If age mattered,

then we should see variability across the age groups, perhaps with a clearer distinction of more plausible (in context) sentences from implausible sentences in the oldest participant group compared with the two younger age groups. And if sentence structure matters and is considered in plausibility ratings, then we should see lower plausibility ratings for sentences in condition 2, i.e., b than 1, i.e., a [see Examples A 1.a) and 1b); B 1.a) and 1.b)].

For both pilot studies, ratings were relatively low, in line with expectations, with the exception of the older adults in the German study (see Figure 2). Context did boost the ratings compared with no context (English Pilot 1: context $M = 3.26$ compared with $M = 3.13$ for no context; German Pilot 2: context $M = 4.64$, and no context, $M = 3.84$). Age also mattered: in the English-language rating simulations, context reversed the sentence structure preference in the oldest participants only. For the younger groups, it merely boosted their ratings. This insight from the English pilot is in line with the view that context can influence comprehension more for older than younger participants. But for the German-language rating simulations, the older adults (55–65 years) rated sentences high no matter the context or the sentence structure; simulations show comparatively lower ratings for both sentence conditions in the 16–18 and 18–31 year olds with context. Somewhat higher ratings were given to the more frequent and easy-to-process sentence types (main clauses, subject–verb–object order) than the less frequent and difficult-to-process sentences

TABLE 4 Detailed analysis of target-context relations with English translations.

Condition	Target sentence (German)	English translation	Plausibility (isolation)	Context overlap	Stereotypical/described competitor in context
1	Den Skifahrer (obj) unterwirft der Zauberkünstler (subj).	The magician subdues the skier.	Implausible	Full event match	Yes: <i>conqueror</i> is a stereotypical agent for <i>subdue</i> and present in context
2	Den Skifahrer (obj) unterwirft der Eroberer (subj).	The conqueror subdues the skier.	Highly plausible, stereotypical	Partial (verb + patient)	Yes: <i>magician subdues skier</i> explicitly given in context
3	Den Skifahrer (obj) unterhält der Eroberer (subj).	The conqueror entertains the skier.	Implausible	Full event match	No competing stereotypical agent in context
4	Den Skifahrer (obj) verwandelt der Zauberkünstler (subj).	The magician transforms the skier.	Highly plausible, stereotypical	Agent-only overlap	No competing stereotypical agent in context

Context: Zauberkünstler unterwirft Skifahrer (“magician subdues skier”); Eroberer unterhält Skifahrer (“conqueror amuses skier”). English translations are phrased in subject–object (instead of German object–subject) constituent order; they are intended to preserve core event structure otherwise.

(reduced relative clauses and object–verb–subject order). This could mean that the grammatical aspects of language influenced plausibility ratings simulated in ChatGPT5 (in line with the definition of plausibility provided by the LLM). This tendency was more pronounced for the English pilot study than for the German study, perhaps reflecting that the LLM captures English-language properties better than German language properties. Alternatively, there is something systematically different in how different age groups assess plausibility across English and German. This could be tested in future experiments with human participants.

3 Main experiments

Recall that the main experiments examined the effects of context presence, age, and context-sentence relation (Experiment 1a), and of genre and context-sentence relations (Experiment 1b) on simulated plausibility ratings (see Section 1 for hypotheses of the between-participant factors). A new set of materials was used, and instead of a within-participants sentence structure manipulation, all sentences were German object-initial sentences. The design contrasted plausibility ratings for four sentence-context relations within participants. This was done to assess whether LLM-simulated plausibility ratings for relatively low-plausibility events in non-canonical object-initial sentences in German can be boosted by means of an event context description.

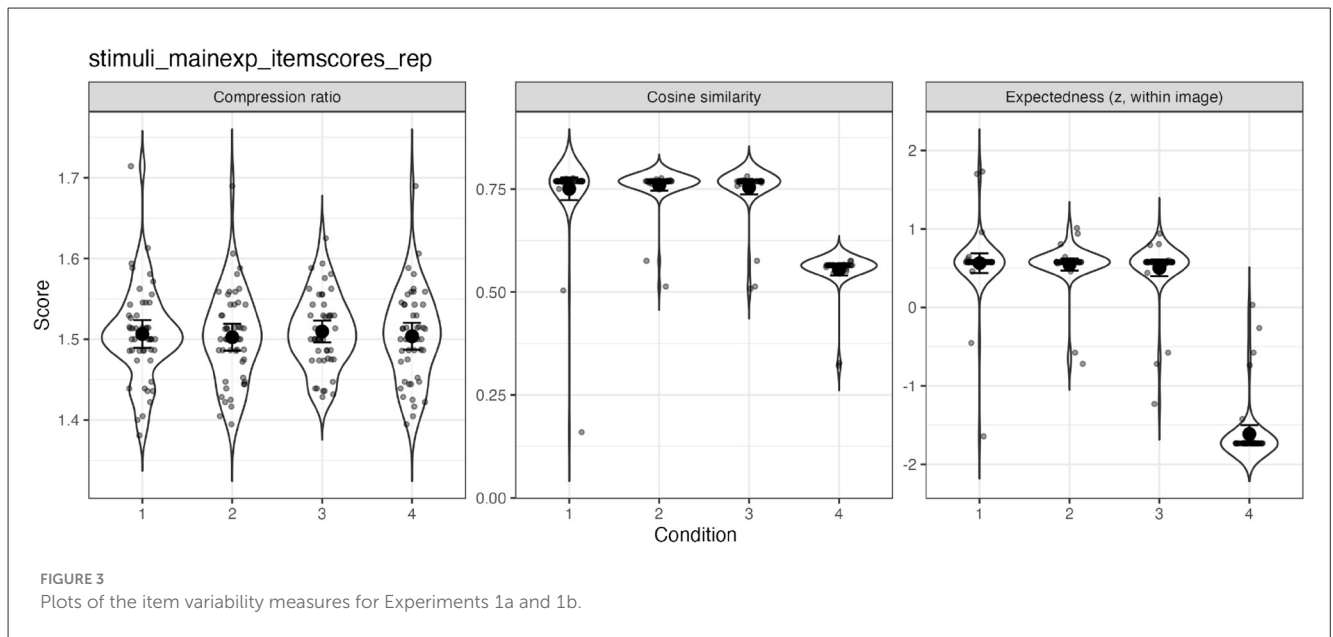
3.1 Materials and design

In Experiment 1a, the event context (context vs. no context), and age (16–18, 18–31, and 55–65) were manipulated between participants. In Experiment 1b, text source (fiction vs. non-fiction) was manipulated between participants. In both of these experiments, sentence-context relation was also manipulated within participants. Table 4 lists an example item illustrating the four context-target sentence relation Conditions 1–4 (table produced using ChatGPT 5.2). The design contained further counterbalancing: A single experimental item contained in total

eight sentences. In addition to the four target sentences and the contexts listed in Table 4, the experimental item contained a further four sentences. For the analysis, counterbalancing and original conditions were both coded as 1–4. For the example in Table 4, the counterbalancing sentences were: 1’: *Den Skifahrer verwandelt der Eroberer*; 2’: *Den Skifahrer verwandelt der Zauberkünstler*; 3’: *Den Skifahrer verhaftet der Zauberkünstler*; 4’: *Den Skifahrer unterwirft der Eroberer* (event contexts: *Zauberkünstler verhaftet Skifahrer*; *Eroberer verwandelt Skifahrer*; translation: *magician arrests skier*; *conqueror enchants skier*). Moreover, counterbalancing ensured that each verb appeared in each condition by repeating each verb across two experimental items. For instance, in a further item, the verb *unterhält* (“entertains” from Condition 3 in Table 4) functions with another stereotypical agent (a clown) and a non-stereotypical agent (an inspector) as part of Conditions 1, 2, and 4’. The entire counterbalancing thus comprised 16 sentences and ensured full counterbalancing of thematic fit across conditions.

Post-hoc, as for the pilot studies, (i) surface structural and semantic aspects of the materials and (ii) the relationship of context description to the target sentence were quantified using the same method as for the pilot studies (see Section 2.1 for the methods description). In the main experiment, the target sentences were all object–verb–subject, and this structural similarity was confirmed by non-significant effects of sentence compression ratio. By contrast, context-target sentence cosine similarity differed strongly across the four conditions in both analyses [replication: $F_{(3, 188)} = 113.94, p < 0.0001$], driven by a pronounced reduction in semantic similarity for Condition 4 relative to Conditions 1–3 (less context overlap, see Table 4). Within-context z -scored similarity and rank measures in the main experiment showed large condition effects, with mixed-effects models exhibiting boundary (singular) fits consistent with near-zero residual item variance once condition was included. Figure 3 plots the results.

Design motivation. What this design instantiates is the possibility to tease apart plausibility from contextual presence while assessing the influence of competing information. For instance, in Conditions 1 and 3, the object-initial sentence is implausible (e.g., *Den Skifahrer unterhält der Eroberer*, paraphrased translation: “The skier is being transformed by the conqueror”; a conqueror is not a stereotypical agent of the verb). In both of these conditions, the



sentence event is present as part of the event context; in addition, a competing agent for a subduing action is present in the event context in Condition 1' (a conqueror but not Condition 3 (there is no stereotypical competitor for entertain). In Conditions 2 and 4, the target sentence can be expected to be judged as relatively more plausible—a conqueror stereotypically submits someone and a magician stereotypically transforms someone. But the sentence event was only partially in the event context (literal translation of event context: “magician subdues skier; conqueror entertains skier”). Condition 2 features competing information (a magician subduing the skier); Condition 4 has an agent-only overlap in the context and no competing stereotypical agent.

3.2 Experiment 1a: methods—Prompts and analysis

For the prompts, data, analysis scripts, and results files see <https://osf.io/u2tqr/overview>. The analysis was conducted using scripts based on the pilot studies but with added *post-hoc* contrasts and modified to include four levels of condition.

The selected model for Experiment 1a was: $PlausibilityRating \sim Condition + age_label + context_label + (1 + Condition|ParticipantID) + ((1|Item) + (0 + Condition|Item)) + Condition : age_label + Condition : context_label + age_label : context_label + Condition : age_label : context_label$.

Since there were issues with singularity (= TRUE for the selected lmer model), a reduced model was also fitted (comparing coefficients and standard errors with the full model) to the same data set. Results were qualitatively similar; this convergence suggests that the fixed-effect conclusions do not depend on the singular random-effects structure of the AIC-selected model.

Additionally, a cumulative-link mixed model was fitted (CLMM; ordinal logistic https://user2021.r-project.org/participation/technical_notes/t186/technote/, retrieved October

28, 2025) with random intercepts for participant and item, using *PlausibilityRating* as an ordered outcome. The CLMM converged and produced large, signed fixed-effect estimates for condition, age, and context, and their interactions. However, standard errors and Wald-type *z/p* statistics could not be computed (summary.clmm: “variance-covariance matrix of the parameters is not defined”), because several contrasts exhibited near-complete separation of the ordinal response (i.e., participants rated some conditions almost uniformly high vs. low). This is expected when effects are strong, and it does not undermine the direction of the fixed-effect estimates: The pattern and relative ordering of the CLMM fixed-effect estimates matched the linear mixed model treating plausibility rating as numeric, supporting the robustness of the results.

3.3 Results

3.3.1 Descriptive results

Figure 4 illustrates the results. With context, higher ratings were given across all age groups when the context overlap full event match, than partial (Conditions 2 and 4). Age differentiated the simulated ratings. The 16–18 year-old participant rating simulation produced the sharpest rating distinctions. The same pattern was visible in the other two age groups but was less pronounced than in the youngest age group. What was also noticeable is that lack of plausibility of the target sentence (Conditions 1 and 3) was made up for by context overlap; simulated ratings were even higher than when the target sentence was plausible but the match with the event context only partial. For the no-context conditions, ratings were overall lower (context: $M = 5.05$; no context: $M = 4.31$), suggesting that, the event context did boost the ratings for these sentences. The confidence intervals were substantially wider when no context was given, suggesting more variability in the plausibility ratings compared to when context was present. In addition, without event

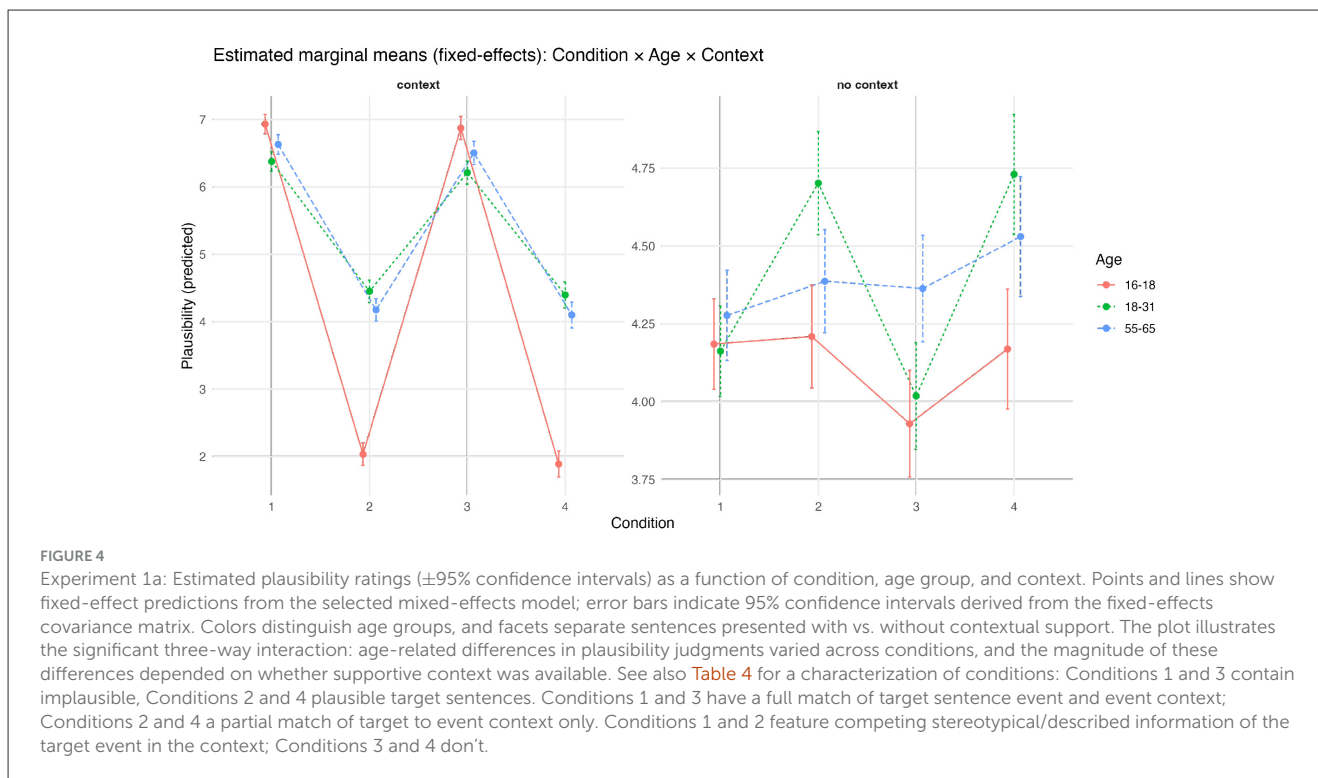


TABLE 5 Experiment 1a: Type-III ANOVA (lmerTest) for the selected model.

Sum Sq	Mean Sq	NumDF	DenDF	F-value	Pr(> F) ^a	Effect
351.907	117.302	3	23.066	188.523	0	Condition
576.846	288.423	2	1,191.417	463.541	0	Age_label
997.532	997.532	1	1,191.417	1,603.188	0	Context_label
3,160.772	526.795	6	2,631.862	846.642	0	Condition:age_label
16,064.986	5,354.995	3	2,631.862	8,606.310	0	Condition:context_label
170.669	85.335	2	1,191.417	137.146	0	Age_label:context_label
2,075.897	345.983	6	2,631.862	556.048	0	Condition:age_label:context_label

^aProbability (Pr) of observing a value greater than the obtained *F*-statistic under the null hypothesis.

context, plausible and stereotypical sentences (Conditions 2 and 4) received higher ratings than implausible sentences (Conditions 1 and 3). This effect was clearest in the 18–31 year olds and somewhat attenuated in the 16–18 and 55–65 year olds.

3.3.2 Inferential analysis

A Type-III ANOVA revealed main effects of condition, age, and context (all p s < 0.001), as well as significant two-way interactions (condition \times age, condition \times context, age \times context) and a robust three-way interaction (condition \times age \times context, p < 0.001, Table 5). This pattern confirms that plausibility ratings varied as a function of all three factors in combination much like in the pilot studies.

The maximal mixed-effects model included random intercepts and slopes for condition by participant and item. The model exhibited a singular fit (one or more random-effect variances near zero). To assess the stability of fixed effects, a simplified model was

re-fitted excluding the random slopes for condition by item. Fixed-effect estimates and standard errors were highly similar across the two models, indicating that the fixed-effect inferences were robust despite singularity in the maximal model. For example, the estimated effect of condition[SC1] was 0.7505 (SE = 0.048) in the full model and 0.7505 (SE = 0.009) in the reduced model.

Post-hoc contrasts: Pairwise comparisons clarified these interactions. **Context-present conditions:** All three age groups differentiated strongly between plausible and implausible sentences. The two implausible sentence conditions were both fully supported by the event context and were rated as more plausible in context than the plausible sentences that were only partially supported by context, irrespective of competitor presence. Thus, participants consistently distinguished event support of sentences but not finer distinctions related to competitor information. For the youngest group (16–18), condition contrasts were very large (e.g., estimate \approx 4.9–5.0, z > 40, p < 0.001). Middle (18–31) and older (55–65) groups also showed significant differences,

TABLE 6 Experiment 1b: Comparison of fixed effects from linear mixed model (LMER) and cumulative link mixed model (CLMM).

Term	LMER (Gaussian)			CLMM (Ordinal, logit)		
	Estimate	SE	<i>p</i>	Estimate	SE	<i>p</i>
Condition [SC1]	0.858	0.060	< 0.001***	2.048	0.041	< 0.001***
Condition [SC2]	-0.126	0.035	< 0.01**	-0.354	0.034	< 0.001***
Condition [SC3]	0.807	0.071	< 0.001***	1.928	0.040	< 0.001***
Genre [SC1]	-0.109	0.023	< 0.001***	-0.190	0.058	< 0.001***
Condition [SC1] × Genre [SC1]	0.435	0.014	< 0.001***	1.068	0.036	< 0.001***
Condition [SC2] × Genre [SC1]	-0.603	0.014	< 0.001***	-1.441	0.037	< 0.001***
Condition [SC3] × Genre [SC1]	0.370	0.014	< 0.001***	0.917	0.036	< 0.001***

Coefficients use sum contrasts for *condition* and *genre*. LMER estimates are in raw plausibility-rating units. CLMM estimates are log-odds from a cumulative logit model; threshold (cutpoint) parameters are not shown. Both models include random intercepts for Participant and Item; the LMER additionally includes random slopes of condition by participant. ****p* < 0.001, ***p* < 0.01.

though with smaller effect sizes (e.g., ≈ 1.9 – 2.0 points, p s < 0.001). In the no-context conditions, most contrasts were attenuated. In younger participants, several condition comparisons were no longer significant (e.g., Condition 1 vs. 2 ≈ -0.02 , $p > 0.1$). Middle-aged and older groups retained some significant contrasts, but effect sizes were modest (≈ 0.4 – 0.6 points). The pattern, however, reversed compared with the context conditions such that plausible (Condition 4) sentences were rated higher than implausible (Condition 3) sentences (e.g., $z = -0.07$, $p < 0.001$ for Condition 4 vs. 3 in 18–31 year olds).

3.3.3 Interim summary

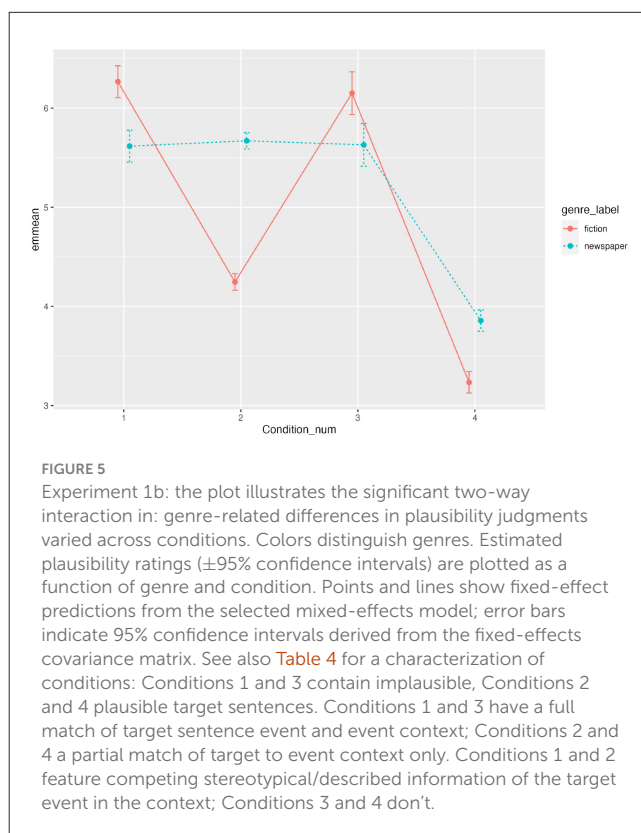
In terms of evaluating the LLM-generated plausibility ratings, plausibility ratings were relatively low when no context was given but boosted overall with context present. When context was absent, the expectation was that plausible sentences (Conditions 2 and 4) would be judged higher than implausible ones (Conditions 1 and 3), and this trend was observed in all three age groups (see Figure 4, no context), validating the ChatGPT-provided ratings. Plausible sentences were penalized and rated lower than implausible sentences when context was present due to less than full event match with the context. Age group mattered with lower ratings for the 16–18 year olds compared to the other age groups.

3.4 Experiment 1b: methods—Prompts and analysis

For the prompts, scripts, results files, and data see <https://osf.io/u2tqr/overview>. The analysis was conducted using the same procedure as for the earlier studies but omitting the factor context presence. The fixed factors were genre and context-sentence relation (four levels as in Experiment 1a).

The selected model for Experiment 1b was: $PlausibilityRating \sim Condition + genre_label + (1 + Condition|ParticipantID) + ((1|Item) + (0 + Condition|Item)) + Condition:genre_label$.

Since there were issues with singularity (= TRUE for the selected lmer model), an ordinal cumulative-link mixed model



(CLMM) fit to the same dataset was also run. Results were qualitatively similar; this convergence suggests that the fixed-effect conclusions do not depend on the singular random-effects structure of the AIC-selected model (see Table 6).

3.5 Experiment 1b: results

3.5.1 Descriptive results

Figure 5 shows a main effect of genre with lower ratings for the fiction ($M \approx 4.97$) than newspaper ($M \approx 5.19$) genre as source for the experiment stimuli. For the fiction genre (red line), ratings were highest in Conditions 1 ($M = 6.27$) and 3 ($M =$

TABLE 7 Experiment 1b: Type-III ANOVA (lmerTest) for the selected model.

Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(> F) ^a	Effect
1053.817	351.272	3	22.490	630.341	0	Condition
12.905	12.905	1	396.920	23.157	0	Genre_label
1,402.258	467.419	3	538.146	838.760	0	Condition:genre_label

^aProbability (Pr) of observing a value greater than the obtained *F*-statistic under the null hypothesis.

6.15), but dropped sharply in Conditions 2 ($M = 4.25$) and 4 ($M = 3.23$). This pattern reflects a clear contextual modulation: Otherwise implausible target sentences received high ratings when the event description supported/named that event too. Relative to that, plausible sentences that were not fully supported in the event description were rated lower. Thus, in fiction, event context exerted a clear influence on plausibility ratings. In contrast, the newspaper genre (turquoise line) showed a relatively flat profile across three conditions ($M \approx 5.5$), except for condition 4 ($M = 3.86$).

3.5.2 Inferential analysis

The selected mixed-effects model included fixed effects of condition, genre, and their interaction, with random intercepts and condition slopes for participants and items. Model singularity was flagged, indicating some overparametrization or redundant variance components, but fixed effects remain interpretable and were corroborated via additional CLMM analysis as reported below.

The ANOVA (Type III see Table 7) and parameter estimates indicate a significant main effect of condition, reflecting that plausibility ratings varied systematically across the four contextual conditions. A significant main effect of genre, whereby texts classified as being from fiction texts were rated as more plausible on average than if the texts were classified as being from newspapers. A significant interaction of condition and genre confirmed that the effect of condition differed by text genre: Context-sentence relations differed in plausibility more for fiction; for newspaper texts, these differences were reduced. Specifically, the presence of the target event in context boosted the ratings for fiction compared with newspaper texts as genre (Conditions 1 and 3); the absence of full overlap decreased the ratings in fiction compared with newspaper texts.

As a robustness check, plausibility ratings were re-analyzed using a cumulative link mixed model (CLMM; ordinal logistic link, ordinal R package), treating the response as ordered categorical (see Table 6). The best-fitting CLMM [PlausibilityOrd ~ condition × genre + (1 | ParticipantID) + (1 | Item)] reproduced the same qualitative pattern as the Gaussian linear mixed-effects model: Significant main effects of condition and genre, and a reliable condition × genre interaction. Including by-participant random slopes for condition did not improve model fit ($\Delta AIC \approx 17$), so the final CLMM included only random intercepts for participant and item. The CLMM confirms that the observed interaction between condition and genre is not an artifact of either the normal-residual assumption or the singular random-effects structure of the LMM.

To further examine the interaction between genre (fiction vs. newspaper) and context-sentence relation (four levels), pairwise

post-hoc contrasts with Holm correction were conducted. In the fiction genre, plausibility ratings differed significantly across most context-sentence relations. Sentences in Condition 1 were rated as substantially more plausible than those in Condition 2, $z = 25.15$, $p < 0.001$, and Condition 4, $z = 36.14$, $p < 0.001$. Ratings in Condition 3 were significantly higher than those in Condition 2, $z = -19.67$, $p < 0.001$. Only one contrast was non-significant, the comparison between the two implausible conditions Conditions 1 and 3, $z = 0.99$, $p = 0.32$.

In the newspaper genre, the pattern of differences was markedly reduced. The only reliable differences emerged for contrasts involving Condition 4. This condition received significantly lower plausibility ratings than Condition 1 ($z = 20.96$, $p < 0.001$), Condition 2 ($z = 40.73$, $p < 0.001$), and Condition 3 ($z = 17.98$, $p < 0.001$). All other contrasts within the newspaper genre were non-significant ($ps \geq 0.42$).

Comparing genres within each condition revealed systematic differences in plausibility ratings. Fiction sentences were rated as significantly more plausible than newspaper sentences in Condition 1 ($z = 13.17$, $p < 0.001$) and Condition 3 ($z = 10.12$, $p < 0.001$). In contrast, fiction sentences were rated as significantly less plausible than newspaper sentences in Condition 2 ($z = -25.49$, $p < 0.001$) and Condition 4 ($z = -11.06$, $p < 0.001$). In summary, ratings simulated with the assumption of a fiction text were generally higher for implausible sentences, whereas ratings simulated with the assumption of a newspaper text as source were higher for plausible sentences.

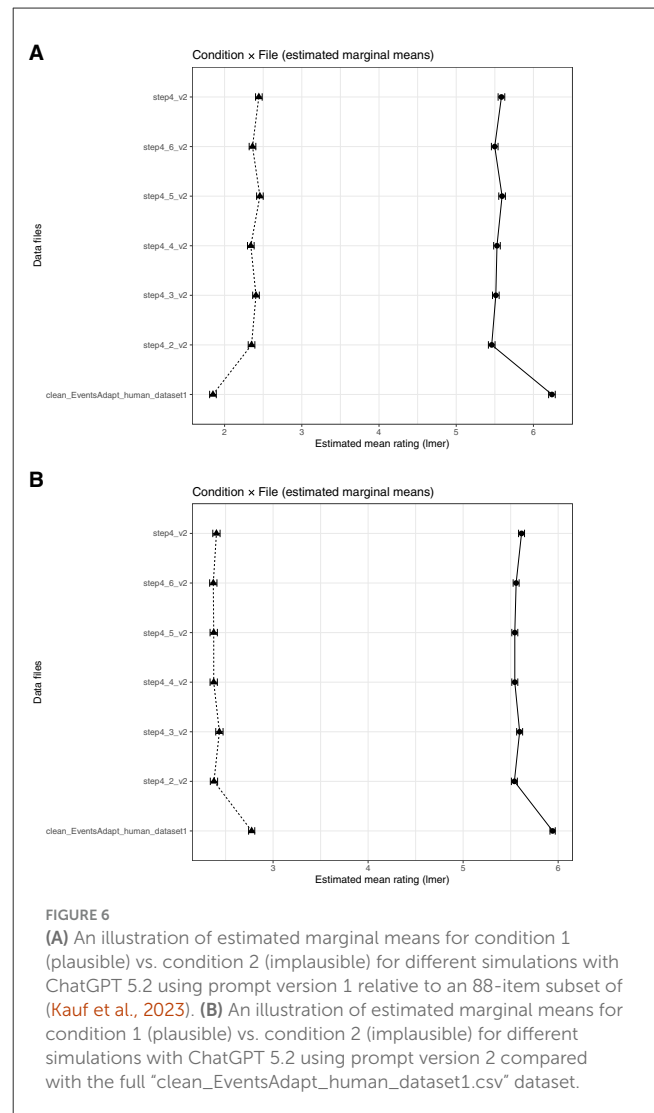
4 Comparison of ChatGPT (5.2) ratings with human plausibility rating data

As a baseline, a direct comparison of LLM-based plausibility rating simulations against some of the human-generated plausibility rating data included in Kauf et al. (2023) was conducted. The same prompting method was used as for the other experiments in the article to obtain the ChatGPT 5.2-based plausibility ratings (see OSF folder “Human_model_comp” for the prompts and further files). First, a subset of “clean_EventsAdapt_human_dataset1.csv, was used $N = 88$ items in ItemNum (filename “Latin_Square_L1_dataset1_AI.xlsx”; 88 AI trials from Kauf et al., 2023 only). In addition, the full data file based on Kauf et al. (2023), labeled “Latin_Square_L1_dataset1.csv” was used. In both cases, the original human ratings were removed, leaving an empty column for “Plausibility Rating,” to be filled with ratings created by the ChatGPT simulation.

Next, prompts (see “prompts_Kaufetal2023_dataset1.docx”) were created. In the file there are three versions of the prompts, each phrased slightly differently to at least somewhat counteract prompt bias. I ran each of these three versions several times: v1 on the subset data file from Kauf et al. (2023), labeled “Latin_Square_L1_dataset1_AI.xlsx”; the other two versions (v2 and v3 prompts) on “Latin_Square_L1_dataset1.csv,” and on the same file with just column headers re-named (e.g., “Condition” for “Plausibility”). All of the files related to these additional runs and analyses can be found on OSF.

These three prompt versions (v1–v3) were run individually (each time opening a new chat for the next run with memory turned off). In addition, a batch version was used (using the first page of the prompt file to ask ChatGPT to run several simulations within one chat, maintaining independence of the ratings as if it were separate experiments). ChatGPT 5.2 was then asked to output the results for these runs in one go. Batch versions v1 and v2 were run twice. For each of these runs with prompt versions v1, v2, and v3, output was obtained as a .csv file with the simulated plausibility rating data. These .csv files were then analyzed and compared with one another, and each .csv file was compared against the human data. For v1 of the prompts, the original human file was reduced to mirror the reduced items in the simulated data. For v2 and v3, the full “clean_EventsAdapt_human_dataset1.csv” file was used.

Figure 6 illustrates the estimated marginal mean of plausibility ratings for the second batch run, with prompt versions 1 and 2. Across all simulations, plausibility ratings showed a robust separation between plausible and implausible conditions. Linear mixed-effects models revealed a strong main effect of condition in every analysis, with plausibility ratings substantially lower in condition 2 than in condition 1 (mixed-effects model fixed-effect for Condition 2: $\beta \approx -4.3$ to -4.4 in versions 1–2 and ≈ -3.2 to -3.3 in version 3; all $ps < 0.001$). For estimated marginal means, all model variants were rated as less plausible on average than the human reference data. Item-level summaries show that the human dataset exhibited stable mean condition differences (v1: 4.39; v2/v3: 3.49). In general, simulations showed smaller differences. In version 1, item-level mean condition differences ranged from approximately 1.3 to 3.1 across the two batch runs and 2.7 to 3.1 in independent-run simulations, yielding deltas relative to the human item-level mean condition difference of roughly $\Delta \approx -1.3$ to -3.0 in the batch runs and $\Delta \approx -1.3$ to -1.7 in independent runs. In version v2, item-level mean condition differences ranged from approximately 3.08 to 4.30 across the two batch runs and 3.27 to 3.36 in independent-run simulations, yielding deltas relative to the human item-level mean condition difference of roughly $\Delta \approx -0.88$ to 0.34 in the batch runs and $\Delta \approx 0.06$ to 0.15 in independent runs. In version v3, item-level mean condition differences ranged from approximately 2.36 to 2.41 across the batch runs and 1.68 to 2.36 in independent-run simulations, yielding deltas relative to the human item-level mean condition difference of roughly $\Delta \approx 1.01$ to 1.06 in the batch run and $\Delta \approx 1.06$ to 1.74 in independent runs. Across versions, condition \times file/run interactions were frequently statistically significant in fixed-effect contrasts, though not always in omnibus Type III Analysis of Variance. Inspection of estimated marginal means and file-specific simple effects indicates that these interactions reflected small quantitative differences in the magnitude of condition differences. What these results suggests



is that the ChatGPT model simulations can reproduce condition differences qualitatively (sentences coded as implausible were given lower ratings than sentences coded as plausible).

5 General discussion

In four pilot studies and two main experiments, the present research assessed whether providing event descriptions can modulate plausibility judgments of target sentences simulated by ChatGPT; whether these context effects on simulated plausibility ratings interact with participant age; and whether they differ by assumed text source/genre (fiction vs. newspaper).

5.1 Effects of context

The present results provided evidence that event context, simulated participant age, and genre can all modulate plausibility judgments simulated by ChatGPT, broadly in line with the hypotheses. This is consistent with prior work showing that

contextual descriptions can activate world knowledge and shift sentence interpretation (Metusalem et al., 2012; Nieuwland and van Berkum, 2006). The analysis of the pilot data and Experiment 1a found that supportive event context boosted the perceived plausibility of target sentences. For Pilot 1, with context, ratings had an average of $M = 3.26$ and were thus higher than in the no context conditions ($M = 3.13$); for Pilot 2, with context, ratings had a mean of $M = 4.64$, and in the absence of event context, $M = 3.84$. In Experiment 1a, implausible sentences were given higher ratings, with context ($M = 5.05$) than without (no context: $M = 4.31$). And within the context-condition, implausible sentences present in the event context had comparable ratings to plausible sentences that were not in the event context (Figure 4). The latter finding broadly held across all three age groups. One interpretation is that the LLM, like human comprehenders, can distinguish plausible from implausible language input and leverage contextual cues to facilitate the interpretation of relatively low-plausibility language input.

5.2 Effects of age group

At the same time, context effects were not uniform across age groups. In Pilot 1, younger adults (16–18 and 18–31) had higher ratings with than without event context, but the 55–65 year olds had slightly lower ratings with event context ($M = 2.58$) than without ($M = 2.85$). In Pilot 2, ratings were lower with context than without for the 16–18 year olds (context: $M = 2.96$; no context: $M = 4.33$). In contrast, higher ratings with context than without emerged in the 18–31 age range (context: $M = 4.51$; no context: $M = 2.38$) and in the 55–65 year olds (context: $M = 6.47$; no context: $M = 4.82$). In Experiment 1a, all three age groups had higher ratings with than without event context, but the size of that difference varied across the age groups. The high ratings for the older groups in Experiment 1a and Pilot 2, and the strong context effects for the 55–65 year olds in Pilot 1 are in line with the hypothesis that the life experience of 55–65 year olds—as encoded in ChatGPT’s knowledge base—renders them likely to rate even implausible events as relatively plausible. Regarding age, a further hypothesis based on Liu (2024) was that semantic plausibility effects might be stronger in older than in younger adults (a bigger difference in ratings between implausible and plausible items), and this is in line with the stronger context boost in the oldest compared to the youngest group Experiment 1a (items with context were simulated as more plausible than on their own). The expectation that there might be no changes in plausibility ratings depending on sentence structure for the older participants was corroborated for German sentences but not for those in English in the pilot studies. Arguably, for the older adults (more than for the other two groups), the greater context-based semantic expectedness of the reduced relative clause than main clause sentences boosted the simulated ratings.

5.3 Effects of sentence structure

The expectation that more difficult sentences elicit lower ratings than canonical and easy-to-process sentences was confirmed

overall in the pilot studies (lower mean ratings for non-canonical than canonical sentences when collapsing across the other factors). The canonical and non-canonical versions of each item were identical except for sentence structure in the German study and were also highly similar in the English study. As a result, these rating differences can be traced back with some confidence to the sentence structure differences. This findings then suggests that ChatGPT—in line with its own definition—includes grammatical coherence (e.g., do sentences follow standard syntactic rules) in its plausibility assessment. These differences were not reflected in measures such as sentence compression ratio, which is supposed to index surface structural similarity but may be reflected in other measures such as dependency scores (total dependency length – larger dependency score for reduced relative than main clauses) or inversion scores (1 if object precedes subject; 0 if subject precedes object).

5.4 Effects of context-sentence relation

For Experiment 1a, younger simulated participants (16–18) showed the sharpest distinction between context-sentence relations when context was present; but their ratings flattened without context. In the absence of context, the variability in the ratings was higher, and the sharpness of distinctions between context-sentence relation conditions decreased, suggesting context had a guiding role when present. Context increased plausibility ratings when the target sentence was implausible but present in the event context (e.g., Condition 1) above the level of plausible sentences for which the event was not in the context (Condition 2). Age interacted with context in shaping how strong these effects were.

5.5 Prior literature

The level of world knowledge in LLMs compared to human world knowledge has been assessed (Kauf et al., 2023). LLMs were good at predicting impossible events but less good at predicting more subtle differentiation between very likely and less likely but possible events. This highlighted a gap between models and humans in how possible and unlikely events are represented. In the meantime—judging by the relatively fine-grained distinctions observed in the present experiments this situation seems to have improved, with model plausibility ratings differentiating implausible from plausible events. At the time, the authors concluded that “a high overlap between the score distributions” was present

for plausible and implausible sentences, meaning that many implausible sentences have higher likelihood generation simply because they contain frequent words.” (Kauf et al., 2023, p. 29)

For the present experiments, this was not the case since the sentences in a condition were fully controlled (they contained the same lexical material); what varied was their relation to an event context (Experiments 1a and b) and their sentence structure (Pilot studies). The studies by Kauf et al. (2023) used different

methods to collect the plausibility ratings and earlier LLM models. It would seem then that by using the graphical user interface with the 2024/2025/early 2026 ChatGPT model, meaningful ratings can be obtained, with perhaps one future use being to derive testable hypotheses of human plausibility ratings in future experiments. This conclusion receives some support from comparisons of Chat GPT 5.2-generated plausibility ratings with human-generated plausibility ratings from Kauf et al. (2023) (see Section 4).

5.6 Caveats and future research

A few caveats are in order. First, prompts can contain biases. Leidinger et al. (2023), for instance, investigated linguistic markers related to mood, tense, aspect, and modality, as well as lexicosemantic variation (e.g., use of synonyms). Their findings revealed that LLMs did not achieve best performance on prompts that reflect language use in pretraining or instruction-tuning data. Further, prompt transfer between datasets or models seemed poor.

They conclude that

“for any model and task, differences in performance can be considerable at even the slightest change in wording or sentence structure.” (Leidinger et al., 2023, p. 9216)

How prompts are phrased can matter, and not reporting prompts makes replication potentially difficult. In the present research, an attempt to counteract prompt bias was made by developing prompts systematically over multiple test runs, and providing the exact prompts on OSF. Furthermore, at least some of the simulations used several different phrasings of prompts (see, e.g., Section 4). That said, in the present simulation experiments, the prompts both described the goal of obtaining plausibility ratings and contained the condition coding; this may have induced bias and reduced variability in the ratings. The condition codes were purely numeric (e.g., “1” and “2”) and did not provide further content. But human participants in an experiment would not be given a clear separation of sentence stimuli into groups via numeric coding and such distinctions are even less obvious given the presence of filler items with a spread of different sentence structures. Future research could mitigate the influence of potential prompt bias by removing the condition coding from the Latin-square input files prior to asking the LLM for the ratings and then adding the condition coding post-rating; in addition, filler items could be included in the rating study and prompts could be varied more strongly and systematically.

A further concern is that the variability inherent in LLM-based ratings is reduced relative to the variability observed across different human participants. The direct comparison of simulated with human ratings from the data by Kauf et al. (2023) confirms this and adds the information that individual runs (starting with a new chat for each run, memory turned off) better approximate human rating variability than batch runs (asking ChatGPT to run several simulations within one chat, maintaining independence of the ratings as if it were separate experiments).

Limitations are also present in the use of LLMs to gain viable pilot data for later experiments with humans via the graphical

user interface. The research on this is in the early stages, and replication as well as generalization are necessary before we can consider this method to be established. It is a potentially useful way of piloting hypotheses. The validity of this approach relies on the LLM successfully distinguishing plausible from implausible stimuli, different sentence-context fits, as well as language use across different age groups and genres. Based on the present study, it seems that at least ChatGPT-4o, 5, and 5.2 can make such distinctions; but which distinctions among the ones reported will resemble human linguistic differentiations (especially concerning the age and genre manipulations) remains to be assessed with human participants on the same materials and designs. The comparison of the LLM-based plausibility ratings obtained via the graphical user interface with human plausibility ratings (Kauf et al., 2023) showed that at least for the basic plausibility distinction, the approach robustly produces the expected higher ratings for plausible compared to implausible stimuli. If one wanted to contrast LLM settings with the effects of context in an experiment, future research could also assess to what extent the prior of an LLM matters in relation to the immediate event contexts. Imagine a prior that is semantically distinct from the target sentences; for this setting, it would be interesting to see to what extent the LLM-simulated ratings would remain sensitive to the event contexts.³

6 Conclusions

A few years ago, capturing world knowledge within a culture and for a specific population such as young healthy or older adults was challenging. But with LLMs, we may have a database from which to assess event context and associated world knowledge when evaluating language. If this assumption holds sufficiently, then simulating world knowledge effects using and LLM’s graphical user interface will help test and refine hypotheses of how world knowledge may affect language processing. With LLMs, we seem to be able to cover a wider range of texts and plausibility ranges, including syntactic ambiguity and picture–sentence incongruence. ChatGPT was able to handle different languages, genres, and age group simulations. Based on the present results, future experiments could assess the goodness of fit of the simulated ratings compared with human ratings of the same sentences and event context sentence combinations across the tested age groups and genres. By establishing a psycholinguistic paradigm for simulating (plausibility-rating) data with ChatGPT-4o and 5, the present research contributes to filling gaps in understanding to which extent empirical paradigms can address theoretical questions, relating computational advances to world knowledge and its variation across age groups and genres.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories

³ I thank a reviewer for this interesting idea.

and accession number(s) can be found at: <https://osf.io/u2tqr/overview>.

Author contributions

PK: Validation, Writing – review & editing, Conceptualization, Writing – original draft, Data curation, Methodology, Formal analysis, Visualization, Investigation.

Funding

The author(s) declared that financial support was received for this work and/or its publication. I gratefully acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), SFB 1412, 416591334 and the individual grants “Effects of lifetime and fact knowledge in comprehension,” KN 897/9-1 and KN 897/9-2. The article processing charge was funded by the Open Access Publication Fund of Humboldt-Universität zu Berlin.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author PK declared that they were an editorial board member of Frontiers at the time of submission. This had no impact on the peer review process and the final decision.

References

- Altmann, G. T. M., and Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition* 73, 247–264. doi: 10.1016/S0010-0277(99)00059-1
- Amsel, B. D., Urbach, T. P., and Kutas, M. (2014). Empirically grounded cognition: the case of color. *NeuroImage* 99, 149–157. doi: 10.1016/j.neuroimage.2014.05.025
- Arai, M., van Gompel, R., and Scheepers, C. (2007). Priming ditransitive structures in comprehension. *Cogn. Psychol.* 54, 218–250. doi: 10.1016/j.cogpsych.2006.07.001
- Atkins, H. G., and Kastner, L. E. (eds.) (1902). *Goethe's Poems*. London: Blackie & Son, Limited.
- Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001
- Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01
- Biber, D., and Conrad, S. (2009). *Register, Genre, and Style*. Cambridge Textbooks in Linguistics. Cambridge, MA: Cambridge University Press. doi: 10.1017/CBO9780511814358
- Brysaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *J. Cogn.* 2, :16. doi: 10.5334/joc.72
- Cai, Z. G., Duan, X., Haslett, D. A., Wang, S., and Pickering, M. J. (2024). Do large language models resemble humans in language use? *arXiv [preprint]*. arXiv:2303.08014 [cs.CL]. doi: 10.48550/arXiv.2303.08014
- Chambers, C. G., Tanenhaus, M. K., Filip, H., and Carlson, G. N. (2002). Circumscribing referential domains during real time language comprehension. *J. Mem. Lang.* 47, 30–49. doi: 10.1006/jmla.2001.2832
- Chen, Q., Zhang, J., Xu, X., Yiming Yang, C. S., and Tanenhaus, M. K. (2016). Prosodic expectations in silent reading: ERP evidence from rhyme

Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. During preparation of this work the author used ChatGPT4o and ChatGPT5 (September 2024–September 2025) to prepare the Open Science Forum pre-registration for which four pilot studies were conducted using ChatGPT4o and ChatGPT5. ChatGPT5 was used to conduct the two main experiments (simulating the data, preparing the analysis scripts, verifying the scripts and interpretation). ChatGPT 5.2 was used to conduct the human-model comparison and the *post-hoc* stimuli analyses (December 2025/January 2026). After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

scheme and semantic congruence in classic Chinese poems. *Cognition* 154, 11–21. doi: 10.1016/j.cognition.2016.05.007

Connell, L., and Keane, M. T. (2006). A model of plausibility. *Cogn. Sci.* 30, 95–120. doi: 10.1207/s15516709cog0000_53

Culler, J. (1975). *Structuralist Poetics: Structuralism, Linguistics and the Study of Literature*. London: Routledge.

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., et al. (2023). Using large language models in psychology. *Nat. Rev. Psychol.* 2, 688–701. doi: 10.1038/s44159-023-00241-5

Early, M. J. (1968). *Teaching Comprehension Skills in Secondary Schools*. Technical report. Washington, DC: ERIC (Education Resources Information Center).

Goulart, L., Gray, B., Staples, S., Black, A., Shelton, A., Biber, D., et al. (2020). Linguistic perspectives on register. *Annu. Rev. Linguist.* 6, 435–455. doi: 10.1146/annurev-linguistics-011718-012644

Guerra, E., Bernotat, J., Carvacho, H., and Bohner, G. (2021). Ladies first: Gender stereotypes drive anticipatory eye-movements during incremental sentence interpretation. *Front. Psychol.* 12:589429. doi: 10.3389/fpsyg.2021.589429

Hagoort, P., Hald, L., Bastiaansen, M., and Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science* 304, 438–441. doi: 10.1126/science.1095455

Hanauer, D. I. (1998). The genre-specific hypothesis of reading: reading poetry and encyclopedic items. *Poetics* 26, 63–80. doi: 10.1016/S0304-422X(98)00011-4

Jackendoff, R. (1997). *The Architecture of the Language Faculty*. Cambridge, MA: MIT Press.

Jakobson, R. (1960). “Linguistics and poetics,” in *Style in Language*, ed. T. A. Sebeok (Cambridge, MA: MIT Press), 350–377.

- Kamide, Y., Scheepers, C., and Altmann, G. T. M. (2003). Integration of syntactic and semantic information in predictive processing: cross-linguistic evidence from German and English. *J. Psycholinguist. Res.* 32, 37–55. doi: 10.1023/A:1021933015362
- Kauf, C., Ivanova, A. A., Rambelli, G., Chersoni, E., She, J. S., Chowdhury, Z., et al. (2023). Event knowledge in large language models: the gap between the impossible and the unlikely. *Cogn. Sci.* 47:e13386. doi: 10.1111/cogs.13386
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. *Psychol. Rev.* 95:163. doi: 10.1037/0033-295X.95.2.163
- Kintsch, W., and Kintsch, E. (2005). "Comprehension," in *Children's Reading Comprehension and Assessment*, eds. S. G. Paris, and S. A. Stahl (London: Routledge), 71–92. doi: 10.1002/9780470757642.ch12
- Knoeflerle, P. (2005). *The Role of Visual Scenes in Spoken Language Comprehension: Evidence From Eye-Tracking* (Doctoral Dissertation in Computational Linguistics). Saarland University. Available online at: <http://scidok.sulb.uni-saarland.de/volltexte/2005/438> (Accessed November 2024).
- Knoeflerle, P. (2025). "Modeling effects of comprehenders' world knowledge on sentence processing," in *Psychology of Learning and Motivation*, eds. K. D. Federmeier, and M. Troyer (Cambridge, MA: Academic Press). doi: 10.1016/bs.plm.2025.07.001
- Knoeflerle, P., and Crocker, M. W. (2006). The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking. *Cogn. Sci.* 30, 481–529. doi: 10.1207/s15516709cog0000_65
- Knoeflerle, P., and Crocker, M. W. (2009). Constituent order and semantic parallelism in on-line comprehension: eye-tracking evidence from German. *Q. J. Exp. Psychol.* 62, 2338–2371. doi: 10.1080/17470210902790070
- Knoeflerle, P., Crocker, M. W., Scheepers, C., and Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: evidence from eye-movements in depicted events. *Cognition* 95, 95–127. doi: 10.1016/j.cognition.2004.03.002
- Kutas, M., and Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event related brain potential (ERP). *Annu. Rev. Psychol.* 62, 621–647. doi: 10.1146/annurev.psych.093008.131123
- Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.* 82, 1–26. doi: 10.18637/jss.v082.i13
- Lau, J. H., Clark, A., and Lappin, S. (2017). Grammaticality, acceptability, and probability: a probabilistic view of linguistic knowledge. *Cogn. Sci.* 41, 1202–1241. doi: 10.1111/cogs.12414
- Leidinger, A., van Rooij, R., and Shutova, E. (2023). "The language of prompting: what linguistic properties make a prompt successful?" in *Findings of the Association for Computational Linguistics: EMNLP 2023*, eds. H. Bouamor, J. Pino, and K. Bali (Singapore: Association for Computational Linguistics), 9210–9232. doi: 10.18653/v1/2023.findings-emnlp.618
- Liu, X. (2024). Age differences in the recruitment of syntactic analysis and semantic plausibility during sentence comprehension. *J. Gen. Psychol.* 151, 444–446. doi: 10.1080/00221309.2023.2283107
- Liu, Y.-T., Nassaji, H., and Tseng, W.-T. (2024). Effects of internal and external attentional manipulations and working memory on second language vocabulary learning. *Lang. Teach. Res.* 28, 1701–1741. doi: 10.1177/13621688211030130
- Matsuki, K., Chow, T., Hare, M., Elman, J. L., Scheepers, C., McRae, K., et al. (2011). Event-based plausibility immediately influences on-line language comprehension. *JEP:LMC* 37, 913–934. doi: 10.1037/a0022964
- McNamara, D. S. (2001). Reading both high-coherence and low-coherence texts: effects of text sequence and prior knowledge. *Can. J. Exp. Psychol.* 5:51. doi: 10.1037/h0087352
- McNamara, D. S., Floyd, R. G., Best, R., and Louwerse, M. (2004). "World knowledge driving young readers' comprehension difficulties," in *Proceedings of the 6th International Conference on Learning Sciences, ICLS '04* (Routledge: New York), 326–333.
- Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., Elman, J., et al. (2012). Generalized event knowledge activation during online sentence comprehension. *J. Mem. Lang.* 66, 545–567. doi: 10.1016/j.jml.2012.01.001
- Nieuwland, M., and van Berkum, J. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *J. Cogn. Neurosci.* 18, 1098–1111. doi: 10.1162/jocn.2006.18.7.1098
- Pescuma, V. N., Serova, D., Lukasek, J., Sauerermann, A., Schäfer, R., Adli, A., et al. (2023). Situating language register across the ages, languages, modalities, and cultural aspects: evidence from complementary methods. *Front. Psychol.* 13:964658. doi: 10.3389/fpsyg.2022.964658
- Pickering, M. J., and Branigan, H. P. (1998). The representation of verbs: evidence from syntactic priming in language production. *J. Mem. Lang.* 39, 633–651. doi: 10.1006/jmla.1998.2592
- Pickering, M. J., and Traxler, M. (1998). Plausibility and the recovery garden paths: an eye-tracking study. *J. Exp. Psychol. Learn. Mem. Cogn.* 24, 940–961. doi: 10.1037/0278-7393.24.4.940
- Plaisance, S. (1928). A reply to Mr. Wald. *Mod. Lang. J.* 12, 651–652.
- Rommers, J., Meyer, A. S., Praamstra, P., and Huettig, F. (2013). The contents of predictions in sentence comprehension: activation of the shape of objects before they are referred to. *Neuropsychologia* 51, 437–447. doi: 10.1016/j.neuropsychologia.2012.12.002
- Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., and Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition* 71, 109–148. doi: 10.1016/S0010-0277(99)00025-6
- Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science* 268, 632–634. doi: 10.1126/science.7777863
- Thornton, R., and MacDonald, M. (1997). Plausibility and grammatical agreement. *J. Mem. Lang.* 48, 740–759. doi: 10.1016/S0749-596X(03)00003-2
- Wertgen, A., and Richter, T. (2020). Source credibility modulates the validation of implausible information. *Mem. Cognit.* 48, 359–1375. doi: 10.3758/s13421-020-01067-9
- Zwaan, R. (1994). Effects of genre expectations on text comprehension. *J. Exp. Psychol. Learn. Mem. Cogn.* 20, 920–933. doi: 10.1037//0278-7393.20.4.920