“Zero-Shot Language Learning” : Can Large Language Models Acquire Contextual Emotion “Like Humans” ?

Authors: Wu Shiyu, Wang Yiyun, Wu Shiyu

Date: 2025-10-17T07:47:27+00:00

Abstract

This study aims to examine whether large language models (LLMs) can incidentally acquire the emotional valence of words from their contextual emotions under “zero-shot” conditions, and to evaluate the effects of emotional valence and contextual variability on vocabulary learning. The study employs a cross-model-human comparison design, where four LLMs and three groups of learners study target words embedded in different emotional contexts (positive/neutral/negative) and repeated/varied contexts using unified materials, with multiple tests measuring affective transfer as well as orthographic and semantic acquisition outcomes. Results demonstrate that LLMs exhibit patterns consistent with humans, transferring contextual emotions to target words and maintaining emotional consistency in language generation; they also show a “positive emotion advantage” and a “contextual variability advantage,” with an interaction effect between contextual emotion and contextual variability emerging in definition generation. The article proposes a “dual-mechanism framework,” suggesting that LLMs possess human-like affective-semantic learning at the functional level, but their mechanisms are based on statistical co-occurrence and vector optimization, distinct from human embodied and social processing. This study provides implications for affective computing, human-computer interaction ethics, and vocabulary instruction.

Full Text

Zero-Shot Language Learning: Can Large Language Models Acquire Contextual Emotion in a Human-Like Manner?

WU Shiyu, WANG Yiyun
School of Foreign Languages, Shanghai Jiao Tong University, Shanghai 200240, China

Abstract

This study investigates whether large language models (LLMs) can incidentally acquire the contextual emotion of words through reading under zero-shot conditions, and evaluates how emotional valence and context variability influence vocabulary learning. Using a cross-model–human comparative design, four LLMs and three groups of learners studied target words embedded in positive, neutral, and negative contexts under repeated versus varied exposure conditions, with multiple tests measuring emotional transfer and acquisition of word form and meaning. Results show that LLMs mirror human patterns: they transfer contextual emotion to target words and maintain emotional consistency in language generation, exhibiting both a "positivity advantage" and a "context variability advantage."Moreover, an interaction between contextual emotion and variability emerged in definition generation.

The paper proposes a "Dual-Mechanism Framework," arguing that LLMs achieve functionally human-like affective-semantic learning through statistical co-occurrence and vector optimization, distinct from human embodied and social processing. These findings offer insights for affective computing, human-AI interaction ethics, and vocabulary pedagogy.

Keywords: large language models, zero-shot learning, contextual emotion, incidental vocabulary acquisition, context variability, positivity advantage

1.1 Zero-Shot Learning and Incidental Vocabulary Acquisition

In human cognitive systems, emotion is not a decorative appendage to rationality but a structural resource (Damasio, 1994). Neuroscientific and cognitive psychological research has established that emotion not only regulates attention and consolidates memory (Eysenck & Brysbaert, 2018) but also deeply shapes value judgments, moral reasoning, and social behavioral choices (Barrett, 2017). In language, culture, and interaction, emotion is not merely "content" for external expression but an internal mechanism that drives communicative intentions, constructs shared perspectives, and coordinates interpersonal consensus. It is a fluid cultural practice: through semantic associations and pragmatic conventions, emotion is continuously endowed with meaning in social interaction and shapes human cognitive structures of the world (Ahmed, 2004, 2010).

Systematic investigation of affective acquisition mechanisms represents a critical pathway to revealing the uniqueness of human cognition. Individuals are not born with complete emotional judgment capacities; rather, they gradually learn to recognize and express emotions through inductive processes involving context, language, and feedback during interaction (Wetherell, 2012). This process exhibits three key characteristics: context dependency, empathy-driven processing, and interactive plasticity. First, emotion is often implicit in semantic-pragmatic structures, nonverbal cues, and intonation, requiring complex contextual construction abilities for its comprehension (Louw, 1993; Nevisi et al., 2018). Second, affective transfer depends on "brain-to-brain coupling" (Hasson et al., 2012): an individual's emotional signals can trigger neural resonance in others, facilitating cross-individual emotional synchronization (Ho et al., 2023). Additionally, emotional representations are not static structures but cognitive products that are continuously constructed and updated through interaction, from infants' imitation of facial emotions to adults' understanding and adjustment of emotional intentions in social contexts (Tamir et al., 2016).

Thus, the pathway of affective acquisition not only characterizes how humans "understand the world" but also defines how we "become socialized mental beings." In this sense, affective learning constitutes an irreplaceable theoretical entry point for understanding human cognition. Any system attempting to simulate human intelligence that cannot access this mechanism will remain at an essential distance from human-like cognition (Binz & Schulz, 2023).

Current large language models (LLMs), such as ChatGPT and Gemini, have demonstrated remarkable capabilities in language generation, reasoning, and question answering, enabling ordinary users to interact with them in natural language and driving the popularization of artificial intelligence. However, their power extends beyond fluent generation and accurate reasoning to the "human-like characteristics" they exhibit: LLMs can identify linguistic patterns in massive corpora and internalize them into model parameters, analogous to humans accumulating experience to form linguistic knowledge networks. They can also invoke this internalized knowledge for reasoning in new tasks, displaying transfer abilities similar to humans.

Yet the key to determining whether AI can achieve "human-likeness" may not lie in linguistic fluency or reasoning accuracy, but in its capacity to acquire and transmit emotion like humans (Ahmed, 2004, 2010; Barrett, 2017). Affective learning concerns not only the recognition of emotional signals but also their transfer to language use and the construction of emotional consensus in interaction (Hasson et al., 2012; Hatfield et al., 1993). Human mechanisms are built upon the joint drivers of embodied experience, contextual induction, and interactive feedback (Barsalou, 2008; Ellis & Wulff, 2015), whereas current LLMs rely solely on statistical learning from linguistic input, lacking genuine embedding of sensory, motivational, and social contexts (Bisk et al., 2020). Notably, however, LLMs have begun to exhibit certain "surprising" emergent abilities in their rapid evolution, most notably the so-called "zero-shot learning": the potential to autonomously evolve reasoning capabilities through pure reinforcement learning processes without any supervised data (Wang et al., 2019). This ability is common in human language learning; for instance, individuals can understand and generate entirely new concepts and produce never-before-heard sentences based on existing linguistic rules and logical reasoning alone (Chomsky, 1957).

LLMs' zero-shot learning provides an important opportunity to examine their affective learning and transmission capabilities. This study approaches the issue through reading-based incidental vocabulary acquisition, comparing LLMs and human learners in their learning and transmission of contextual emotion, because LLMs' "zero-shot language learning" shares significant similarities with a key mechanism of human language development—reading-based incidental vocabulary acquisition. These similarities manifest on two levels: first, both exhibit non-deliberative learning driven by tasks. Humans infer word meanings through context during discourse reading (Hulstijn, 2001), while LLMs rely on input context and pre-trained knowledge for generation without explicit supervision (Brown et al., 2020). Second, both possess knowledge integration and transfer capabilities. Humans incorporate new words into their mental lexicon through repeated exposure, while LLMs can generalize learned linguistic patterns to entirely new contexts, achieving human-like cross-task output. This mechanistic similarity makes incidental vocabulary acquisition a critical paradigm for evaluating LLMs' affective induction capabilities.

Comparing LLMs and human learners in contextual emotion learning during incidental vocabulary acquisition not only helps clarify whether AI systems possess "human-like" affective learning abilities but also reveals the psychological reality of several core variables in human vocabulary learning mechanisms. Based on this, the present study constructs a cross-human–model comparative paradigm using unified experimental materials and task structures to explore whether contextual emotion in discourse can be induced, transferred, and applied to lexical meaning construction during vocabulary learning under zero-shot conditions. The following sections first review relevant literature on contextual emotion learning and transmission, introduce major theories and empirical findings, identify existing controversies, and finally propose specific research questions.

1.2 Learning and Transmission of Contextual Emotion

Research on contextual emotion learning originated in the 1980s with computer-based corpus linguistics and later expanded into cognitive psychology and applied linguistics, successively proposing the "Contagion Hypothesis," "Transfer Hypothesis," and "Double-Jujube Tree Effect Hypothesis."

The "Contagion Hypothesis" represents the earliest mechanistic theory in contextual emotion learning, originating from Sinclair' s (1987) observation that certain words (e.g., *set in*) frequently appear in negative emotional contexts (e.g., *rot*, *decay*, *disillusion*). Louw (1993) built upon this to propose the concept of "semantic prosody," noting that words acquire stable emotional coloring through frequent co-occurrence with specific affective words. For example, *spark* often co-occurs with positive contexts (e.g., *lively*, *justice*), while *trigger* associates more with negative contexts (e.g., *sudden*, *bankruptcy*, *crash*). Although semantic prosody is not directly reflected in lexical meaning, it profoundly influences semantic processing and emotional judgment, a subtle and nearly implicit semantic phenomenon confirmed by extensive corpus evidence (Nevisi et al., 2018). However, the "Contagion Hypothesis" is primarily based on co-occurrence statistics of existing, semantically stable words in dictionaries and thus cannot explain the "contagion" pathway of emotional coloring during initial encounters with novel words or reveal how contextual emotion migrates to learners' lexical representations.

Snefjella et al. (2020) consequently proposed the "Transfer Hypothesis"of contextual emotion learning. Through variable manipulation and psychological process measurement, they conducted five experiments in which English native speakers read multiple short texts carrying positive, negative, and neutral emotions and learned embedded target pseudowords (see Section 3.2). Results showed that although all target pseudowords expressed neutral semantics in the texts, participants judged them as emotionally positive when encountered consistently in positive contexts and as negative when in negative contexts. This confirmed that contextual emotion successfully transferred from context to learners' semantic representations of target words and became internalized as part of their lexical knowledge. However, despite providing new theoretical and methodological perspectives, the "Transfer Hypothesis" awaits more refined validation of its underlying mechanisms, with potential moderating factors such as learner characteristics, context variability, and exposure frequency requiring further investigation.

Building on this, Wu and Li (2024) proposed the "Double-Jujube Tree Effect Hypothesis." They replicated Snefjella et al.' s (2020) experiments with Chinese native speakers but additionally manipulated context variability (multiple texts vs. multiple readings, see below), having each participant either read the same material multiple times or read multiple different materials carrying the same emotional load before taking vocabulary tests. Results confirmed the transferability of contextual emotion but revealed that such transfer was conditional: it only occurred when the same material was read repeatedly multiple times, manifesting a "repetition-infection" pattern reminiscent of Lu Xun' s "There is another jujube tree" in *Autumn Night*.

Recently, researchers (Li, 2024; Ma & Li, 2024) examined whether L2 learners can also effectively acquire the emotion of contexts in which L2 target words appear, using Snefjella et al.' s (2020) and Wu and Li' s (2024) experimental paradigms in L2 contexts. Findings showed that contextual emotion can transfer from context to target words during L2 reading, but this transfer is moderated by both context variability and learners' L2 proficiency.

In summary, existing research has systematically revealed that human learners can acquire the contextual emotion of target words—semantic prosody—during reading-based incidental vocabulary acquisition. However, the propagation effect of affective learning, that is, whether emotional learning further promotes the acquisition of vocabulary itself (including form and meaning), remains controversial. On one hand, some studies support a "positivity advantage" : compared to neutral or negative contexts, positive contexts better facilitate vocabulary acquisition, manifested as higher recognition rates, stronger semantic integration, and more durable retention (Snefjella et al., 2020). On the other hand, other research proposes a "negativity bias" : negative contexts, due to stronger emotional arousal and cognitive load, confer greater advantages in vocabulary retention (Driver, 2022).

Another core controversy centers on the role of "context variability" (Bolger et al., 2008), namely whether "multiple texts" or "multiple readings" better promotes vocabulary learning (Wu & Li, 2024). According to the Instance-Based Framework, repeated occurrences of words in diverse contexts generate more independent semantic traces, promoting abstraction of core meaning. Experiments by Bolger et al. (2008) and Lauro et al. (2020) support this view, finding that "multiple texts" contexts significantly improve performance in meaning generation and semantic integration. However, other studies have questioned the advantage of varied contexts. Horst et al. (2011) found that repeated contexts ("multiple readings") 反而更有助于词义的学习和保持; Joseph and Nation (2018) also noted that context diversity showed no significant advantage when learning low-frequency words; additionally, Balass (2011) observed results where repeated contexts outperformed varied contexts in some semantic judgment tasks.

Against this backdrop, the present study employs a "zero-shot vocabulary learning"framework to compare LLMs and human learners in affective learning during incidental vocabulary acquisition. This approach can not only clarify whether AI systems possess "human-like" affective learning capabilities but also reverse-test the psychological reality of two key variables in human vocabulary acquisition mechanisms—emotional effects and context variability. Specifically, this study addresses three questions: (1) Can LLMs "learn contextual emotion like humans"? (2) Which context condition, varied or repeated, better promotes LLM vocabulary learning? (3) Does contextual emotion significantly impact vocabulary learning for both humans and LLMs?

2. Method

The experiment consisted of two phases: reading/learning and vocabulary testing. It should be noted that "zero-shot language learning" in this study does not refer to LLMs completing tasks with absolutely no input, but rather that models received no specific supervision or fine-tuning regarding test tasks or target lexical items. They produced immediate output judgments in tests based solely on processing of original reading materials, consistent with the broad definition of current "zero-shot" tasks (Brown et al., 2020).

The study aimed to compare the affective learning performance and its impact between LLMs and human learners. Human data were derived from three learner groups: (1) English native speakers from Snefjella et al.' s (2020) experiment; (2) Chinese native speakers from Wu and Li' s (2024) experiment; (3) English L2 learners from the authors' recent experiment (complete data to be reported separately). All three learner groups used consistent experimental tasks and structures, with materials differing slightly by language: English native speakers and L2 learners used Snefjella et al.' s (2020) original English materials, while Chinese native speakers used a rigorously semantically equivalent Chinese version to ensure structural and contextual consistency. For lexical item construction, the English experiments used pseudowords generated by Wuggy, while the Chinese experiment followed the same construction principles, using phonetic transliterations from Arabic as a basis for phonological and orthographic modification to create natural-looking but semantically unloaded Chinese pseudowords. For convenience, these three experiments are collectively referred to as "baseline experiments."

2.1 Large Language Models

This study selected four representative LLMs: Ernie Bot 3.5, ChatGPT (GPT-4), Gemini (1.5 Pro), and LLaMA (3.1-8B). Model selection followed these principles (see Table 1): (1) covering current mainstream technical architectures, including standard Transformer architectures (e.g., LLaMA, ChatGPT) and hybrid architectures integrating MoE mechanisms (e.g., Gemini); (2) reflecting diverse training backgrounds and language adaptability, covering bilingual Chinese-English optimized models (e.g., Ernie Bot as Chinese-optimized) and English-dominant models (e.g., ChatGPT and Gemini); (3) encompassing different scales and open-source characteristics, including both closed-source and open-source models. These models differ significantly in architectural design, parameter scale, and context window length, while also demonstrating complementarity in text emotion processing, semantic generation, and contextual adaptation capabilities, providing strong comparative value.

Table 1 Introduction to the Four Large Language Models

Model	Context Length (tokens)	Architecture & Parameters	Multimodal Reasoning Capability
Ernie Bot 3.5	-	Transformer/—	Supports text, image, and video generation; moderate reasoning capability, suitable for Chinese contexts
ChatGPT-4	1.75 trillion	Transformer/—	Strong image generation and mathematical/logical reasoning; supports text-image integration
Gemini 1.5 Pro	-	Transformer, MoE/ —	Strong visual understanding and generation; powerful complex question and knowledge query reasoning
LLaMA 3.1-8B	-	Transformer/ 8B	Text generation, lacks multimodal capability; moderate-scale tasks, basic reasoning

It should be noted that these four models represent different technical paths and language capability optimization directions. They are not absolutely comparable in performance but provide an explanatory comparative foundation across key dimensions such as architecture, openness, language adaptability, and reasoning capability. Therefore, under unified input formats, task types, and output evaluation, this study conducted parallel testing of the four models' performance to explore different processing pathways and behavioral patterns in contextual emotion acquisition tasks. Model outputs and the three human learner groups together form seven comparable datasets (4 LLMs + 3 learner groups), providing a solid foundation for examining the human-likeness and boundaries of contextual emotion learning and semantic prosody transfer mechanisms.

2.2 Materials

For convenient comparison, LLM learning materials were identical to the first and third baseline experiments, including: reading passages, target words embedded in passages, and vocabulary knowledge test items. See Snefjella et al. (2020) and Wu and Li (2024) for details.

2.2.1 Target Words Nine target words were examined, all pseudowords generated by Wuggy software (Keuleers & Brysbaert, 2010) used in Snefjella et al. (2020): *aunith, ceammy, cruce, flyph, mernt, neak, plurk, rotch*, and *wurge*. These pseudowords are orthographically and phonologically similar to real English words, conforming to English orthographic and phonological rules.

Each target pseudoword was matched with five additional pseudowords: one homophonic pseudoword (e.g., *crooce* for *cruce*) served as a distractor in the orthographic choice task, while four non-homophonic pseudowords served as distractors for the other four vocabulary knowledge tests (see below), generating 45 distractor pseudowords total. All pseudowords underwent standardized correction to ensure no significant differences in emotional valence between target and filler pseudowords (see Snefjella et al., 2020; Wu & Li, 2024).

2.2.2 Reading Materials Reading materials were designed in triplets, with target words embedded in each triplet. Each triplet contained three different short texts, as in example (1):

(1) a. Her wonderful teacher performed at the exciting concert.
She was playing the (NONWORD).

b. Her new teacher performed at the normal concert.
She was playing the (NONWORD).

c. Her annoying teacher performed at the horrible concert.
She was playing the (NONWORD).

Texts in each triplet contained two sentences. The first sentence manipulated contextual emotion through adjectives and nouns at three levels (positive, neutral, negative), while the second sentence was identical across the triplet, containing the target word to be learned. The context formed by the first and second sentences ensured learners could infer each target word' s meaning, i.e., its denotation. All target words expressed neutral emotional denotations and represented common everyday items (broad semantic categories), such as a musical instrument, plant, or kitchen utensil. Example (1) represents a musical instrument. There were nine denotations total.

Each target word was distributed across five different triplets, resulting in 45 triplets (9 × 5) and 135 short texts (45 × 3). Each learner read 45 texts in the experiment, with each target word appearing five times total, ensuring vocabulary acquisition through reading (e.g., Blythe et al., 2012; Godfroid et al., 2018). Reading materials were generated using templates pre-designed by Snefjella et al. (2020) to control for potential confounding variables such as context length and target word frequency.

Contextual emotion (positive, neutral, negative) was a within-subjects variable. Using a Latin square design, the 135 reading materials (45 triplets) were cross-balanced across the three emotion levels and nine denotations, generating nine material sets. In each set, every target word appeared in five different texts, always in the same emotional context and expressing the same denotation. Through the nine sets, each target word appeared in every emotion level and expressed every denotation, achieving comprehensive and balanced coverage. Additionally, to test the context variability hypothesis, context variability (multiple texts vs. multiple readings) was manipulated as a between-subjects variable: for each target word, participants either read the same text five times or read five different texts. This resulted in 18 experimental material sets (9 × 2). This design aimed to avoid attention dispersion and memory load increases due to excessive task demands while maintaining consistency with prior baseline studies to enhance comparability; mixed-effects models were used in statistical analysis to control for variation caused by individual differences.

2.2.3 Vocabulary Test Items Five vocabulary knowledge tests were designed to comprehensively evaluate LLM learning outcomes from different dimensions: emotional valence rating task, sentence production task, orthographic choice task, definition matching task, and definition generation task. Except for the sentence production task, the other four tasks were identical to baseline experiments.

The **emotional valence rating task** tested LLMs' learning of contextual emotion. It included 18 test words: nine target words (e.g., *flyph*) and nine filler words (e.g., *snicle*). Words were presented in pseudorandom order. LLMs first judged whether they had seen the word in the preceding reading; if "no," they skipped to the next word; if "yes," they rated its emotion (1 = "sad," 9 = "pleasant").

The **sentence production task** also tested LLMs' learning of contextual emotion. Eighteen words were presented in pseudorandom order. LLMs first judged whether they had seen the word; if "no," they skipped it; if "yes," they generated a sentence using the word (maximum 30 words). Nine were target words, nine were fillers. This task was added beyond baseline experiments because language generation is a core LLM capability. Through sentence production, we could evaluate LLMs' contextual emotion learning in scenarios closer to real language use, better reflecting model performance in practical applications.

The **orthographic choice task** measured LLMs' mastery of target word forms. LLMs judged whether 18 words had appeared in the reading: nine target words (e.g., *flyph*) and nine filler pseudowords (e.g., *fliph*). The ability to distinguish target words from fillers (homophones) measured memory for word form.

The **definition matching task** assessed LLMs' meaning recognition. LLMs selected the definition matching each word from given options, with 18 test words (nine targets, nine fillers). The **definition generation task** tested meaning recall. With 18 test words (nine targets, nine fillers), LLMs first judged whether they had encountered the word; if "no," they skipped it; if "yes," they generated its definition, including Chinese translation, English paraphrase, or synonym.

Although both definition matching and generation tested semantic learning, generation requires retrieving meaning from memory, making it more difficult than recognition (selecting from options) and assessing more advanced lexical-semantic knowledge (Laufer & Aviad-Levitzky, 2017; Stewart et al., 2024).

2.3 Procedure

This study used data from the third baseline experiment (English L2 learners) as a blueprint, conducting LLM simulation and data collection on a one-to-one participant basis. The baseline experiment included 306 Chinese English learners, each reading 45 texts and completing four vocabulary tests. Accordingly, we constructed a corresponding LLM testing flow for each human participant: each independent testing session was treated as a unique observation unit, functionally equivalent to "one participant" in human experiments. Therefore, each LLM (four total) completed 306 sessions independently, generating 1,224 "model participants" total. In each session, reading material presentation order strictly replicated that of its corresponding human participant, ensuring comparability with human data while preserving randomness in material presentation.

To effectively ensure inter-session independence, this study employed strict session isolation protocols: each session launched in a fresh dialogue window where the model sequentially completed reading of 45 texts and subsequent five tests (including sentence production); once the session ended, dialogue context and history cache were completely cleared before opening a new session, preventing any carryover of information from previous sessions. This procedure effectively prevented cross-session "information contamination,"satisfying core assumptions of mixed-effects models (see below) regarding observation independence and ensuring comparability and interpretability across different LLMs. All tests were completed under zero-shot conditions; prompts (see appendix) controlled only task flow and response format, providing no examples, cues, or training inputs to avoid information leakage and ensure reproducibility and verifiability.

In terms of task flow, after LLMs completed reading 45 texts in each testing session (representing one participant), they immediately underwent five vocabulary tests in this order: orthographic choice task, emotional valence rating task, definition generation task, definition matching task, and sentence production task. Throughout the process, models received no examples or feedback, ensuring all responses were generated strictly under zero-shot conditions.

3.1 Contextual Emotion Acquisition

LLMs' contextual emotion acquisition is reported from the word emotional valence rating task and sentence production task. Since original English native speaker data were unavailable, Snefjella et al.' s (2020:8-13) reported results are summarized. For Chinese native speakers and L2 learners, their data were merged with LLM data for comparative analysis.

3.1.1 Emotional Valence Rating

Snefjella et al. (2020:8-13) found that English native speakers' emotional valence ratings for target words in positive contexts were always highest, while ratings for words in negative contexts were always lowest. Statistical inference results further confirmed significant differences in ratings across the three contextual emotion conditions (negative vs. neutral vs. positive), verifying successful transfer of contextual emotion to target words.

Figure 1

presents contextual emotion learning results for the two learner groups (Chinese native speakers, L2 learners) and various LLMs.

Figure 1 Emotional Learning Results for Language Learners and LLMs

Figure 1 shows that the two learner groups exhibited very similar overall patterns in emotional valence ratings. More importantly, each LLM—whether Ernie Bot, ChatGPT, Gemini, or LLaMA—showed rating patterns very similar to both learner groups: ratings were always highest for positive contexts and lowest for negative contexts.

Using the `clmm()` function from R' s `ordinal` package, cumulative link mixed-effects models (Christensen, 2023) were fitted to emotional valence ratings (ordinal data) for both learner groups and each LLM. Fixed effects included contextual emotion (negative, neutral, positive) and learner type (Chinese native speakers, L2 learners, and each LLM). Random effects included participant (or model-simulated participant), vocabulary test item, and target word' s denotation. Model fitting results showed that regardless of learner type—whether the

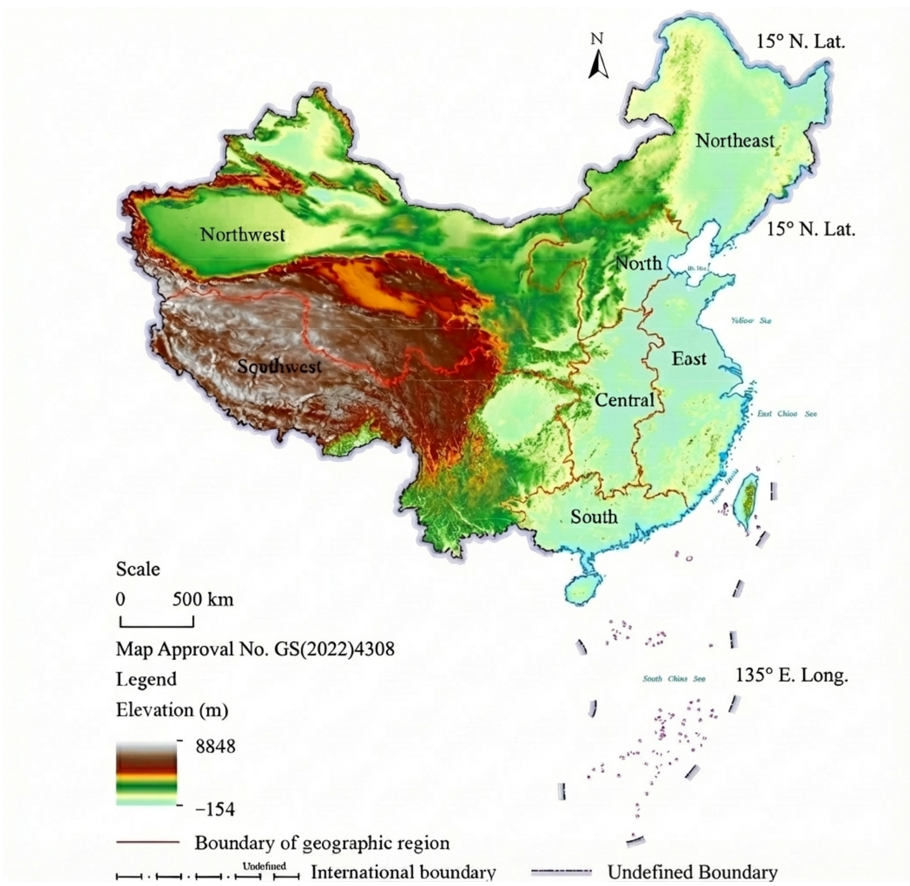

Figure 1: Figure 1

two learner groups (L1 and L2) or various LLMs—emotional valence ratings for target words across the three contextual emotion conditions differed significantly (zs < -9.00, ps < .0001), showing the pattern: positive > neutral > negative. As noted earlier, since these words were pseudowords without inherent meaning and expressed only neutral, concrete, broad semantic categories (e.g., a ship, a kitchen utensil) when embedded in texts, the only possible source of differences in participants' post-reading emotional valence ratings was the context—they learned the words' emotion from the contexts in which they appeared.

3.1.2 Sentence Production Emotion Analysis

All language models achieved high accuracy in sentence production, exceeding 90% except for Ernie Bot at 89.10%. The Python natural language processing tool TextBlob was used to conduct sentiment analysis on LLMs'correct sentences. TextBlob generates sentiment scores ranging from -1 to 1 for each sentence, where negative values indicate negative emotion, positive values indicate positive emotion, and higher scores indicate stronger positive emotion. Figure 2

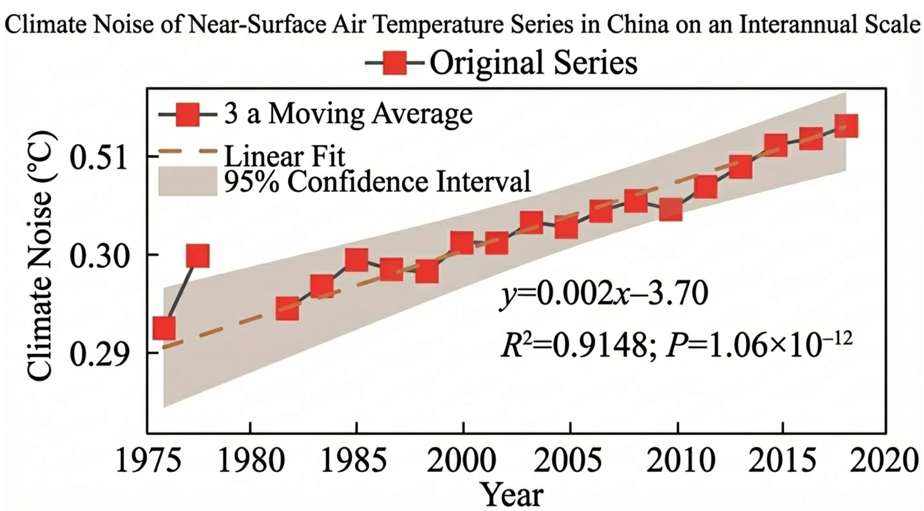

Figure 2: Figure 2

presents the analysis results.

The overall pattern in Figure 2 closely resembles Figure 1. When LLMs used target words learned in positive contexts, generated sentences always had the highest sentiment values, while words from negative contexts produced sentences with the lowest values. More specifically, LLaMA showed the most pronounced emotional learning effect, followed by ChatGPT, then Gemini, and finally Ernie Bot.

Figure 2 Sentiment Analysis Results for Sentences Produced by LLMs

Using the `lmer()` function from R's `lme4` package, mixed-effects models (Baayen et al., 2008) were fitted to sentiment values of sentences produced by each LLM. Fixed effects included contextual emotion (negative, neutral, positive) and LLM type. Random effects included model-simulated participant, vocabulary test item, and target word's denotation. Model fitting results showed that except for Ernie Bot, the other three LLMs produced sentences with significantly different sentiment values across the three contextual emotion conditions (zs < -2.14, ps < .05), showing the pattern: positive > neutral > negative. For Ernie Bot, although no significant difference existed between positive and neutral (β = -0.004, SE = 0.01, z = -0.29, p = .95), differences between neutral and negative and between positive and negative were significant (zs < -7.10, ps < .0001).

These results demonstrate that despite individual differences among LLMs, they all successfully learned the contextual emotion of words—semantic prosody—from the perspective of language production.

3.2 Word Form and Meaning Acquisition

Since Snefjella et al.' s (2020) experiment did not examine context variability effects, only the second and third baseline experiments are compared here: two learner groups (Chinese native speakers, L2 learners) and LLMs. Word form and meaning acquisition were measured through three test tasks: orthographic choice (form knowledge), definition matching (meaning recognition), and definition generation (meaning recall). All test task answers were coded as "correct" or "incorrect" —for example, if learners or LLMs selected the correct word form, correctly matched a definition, or provided a correct definition, the answer was coded as "correct," otherwise "incorrect."

The `glmer()` function from the `lme4` package was used to fit logistic regression mixed-effects models to learner and LLM performance on these three tasks. The model had two independent variables: context (repeated vs. varied) and contextual emotion (negative vs. neutral vs. positive). However, contextual emotion was not entered as a categorical variable; instead, the emotional valence ratings that learners or LLMs gave each target word were entered into the model, allowing more intuitive assessment of contextual emotion' s (learning effect) impact (see Snefjella et al., 2020). This variable was standardized before model entry to facilitate fitting and interpretation. Random effects included participant (learner or model-simulated learner), vocabulary test item, and target word' s denotation.

3.2.1 Orthographic Choice Task

Chinese native speakers and L2 learners achieved mean accuracies of 89.17% and 82.65%, respectively. However, except for LLaMA (98.8%), all LLMs achieved perfect accuracy (1.0), demonstrating stronger word form learning capabilities.

3.2.2 Definition Matching Task

Figure 3

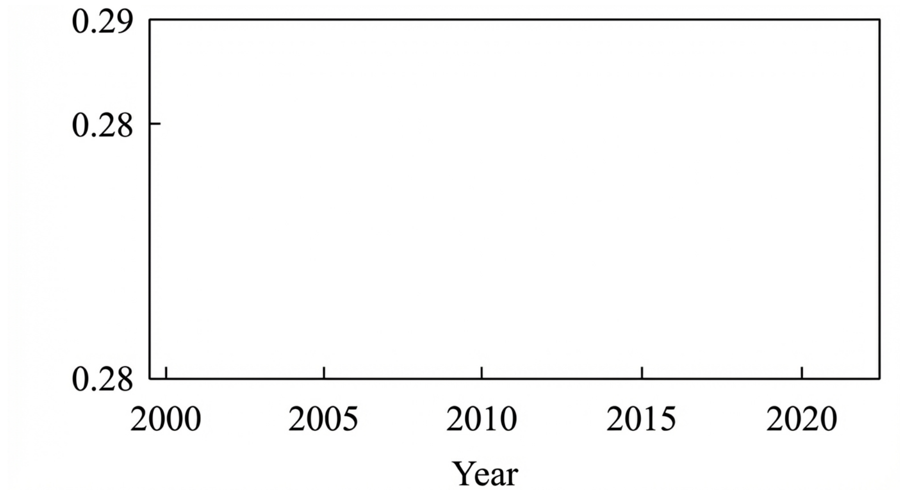

Figure 3: Figure 3

presents mean accuracy for learner groups and each LLM on the definition matching task under different context conditions (repeated vs. varied).

Figure 3 Mean Accuracy on Definition Matching Task for Language Learners and LLMs in Repeated and Varied Contexts

Figure 3 shows that LLMs' learning effects were generally better than language learners' , confirmed by mixed-effects model results (β = 1.88, SE = 0.07, z = 24.45, p < 0.001). However, LLMs also showed a pattern highly consistent with learners: learning effects were always better under "multiple texts" (varied) contexts than "multiple readings" (repeated) contexts, also confirmed by mixed-effects model results (β = 1.13, SE = 0.21, z = 5.720, p < 0.001).

Given the consistency of these variable effects, for convenient comparison, data from both learner groups and each LLM were first merged separately before constructing mixed-effects models for statistical analysis. Results showed that context variability and contextual emotional valence had very similar effects on both experimental groups: (1) **Positivity advantage**: as contextual emotional valence increased, meaning recognition ability improved significantly for both learner groups and each LLM at nearly identical magnitudes (0.21 vs. 0.23) (zs > 7.64; ps < 0.001). (2) **Varied context advantage**: varied contexts produced significantly better learning effects than repeated contexts for both learner groups and each LLM (learners: β = 0.71, SE = 0.28, z = 2.56, p = 0.01; LLMs: β = 2.02, SE = 0.25, z = 7.94, p < 0.001). (3) No interaction existed between contextual emotional valence and context variability for either learner groups or LLMs (ps > 0.15), indicating that contextual emotion effects did not depend on context variability levels.

It should be noted that although LLMs and the two learner groups showed highly similar overall trends, they also exhibited some specific differences. For instance, LLMs' overall learning effects were better than both learner groups (significantly higher accuracy, Figure 3). Additionally, although both showed varied context advantages, LLMs' advantage was more pronounced than learners' (effect size: 2.02 vs. 0.71). Regarding individual model performance, Gemini' s meaning recognition was significantly better than all other models (ps < 0.001), while no significant differences existed among the other models (ps > 0.1).

3.2.3 Definition Generation Task

Figure 4 [FIGURE:4] presents mean accuracy for the two learner groups and each LLM on the definition generation task under the two context conditions (repeated vs. varied).

Figure 4 Mean Accuracy on Definition Generation Task for Language Learners and LLMs in Repeated and Varied Contexts

Similar to the definition matching task, LLMs' learning effects were generally better than both learner groups (higher accuracy) ($\beta = 2.48$, SE = 0.08, z = 31.35, $p < 0.001$). However, LLMs still showed a pattern highly consistent with learners: learning effects were always better under "multiple texts" (varied) contexts than "multiple readings" (repeated) contexts ($\beta = 1.29$, SE = 0.21, z = 6.11, $p < 0.0001$).

However, for both learners and LLMs, a significant interaction existed between contextual emotional valence and context variability (learners: $^2(1) = 12.51$, $p < 0.001$; LLMs: $^2(1) = 19.15$, $p < 0.001$), indicating that contextual emotion effects depended on context variability levels (repeated vs. varied). Figure 5 [FIGURE:5] presents the interaction effects of these two variables for learners and LLMs on the definition generation task.

Figure 5 Interaction Effects of Contextual Emotional Valence and Context Variability on Definition Generation for Learners and LLMs

Figure 5 shows that the effects of contextual emotional valence and context variability on learners' and LLMs' performance on the definition generation task (advanced meaning learning) were very similar in overall trend: (1) **Clear varied context advantage**: both learners and LLMs showed better advanced meaning learning under varied contexts than repeated contexts (learners: $\beta = 0.85$, SE = 0.26, z = 3.30, $p < 0.001$; LLMs: $\beta = 2.89$, SE = 0.27, z = 10.80, $p < 0.001$). (2) **Significantly stronger positivity advantage under varied contexts**: under varied contexts, the advantage in advanced meaning acquisition brought by higher contextual emotional valence was significantly stronger than under repeated contexts (learners: $\beta = 0.26$, SE = 0.04, z = 6.96, $p < 0.001$; LLMs: $\beta = 0.56$, SE = 0.08, z = 7.43, $p < 0.001$). This led to (3) **Amplified positivity advantage**: under varied contexts, as emotional valence increased, learners' and LLMs' performance improved further, widening the gap with repeated contexts.

Again, despite high overall similarity between LLMs and the two learner groups, some specific differences emerged. For instance, LLMs' overall learning effects were better than both learner groups (Figure 5). Additionally, although both showed varied context advantages, LLMs' advantage was more pronounced than learners'across both context types. Regarding individual model performance, under varied contexts, all models performed better than Ernie Bot ($ps < 0.02$). Under repeated contexts, Gemini performed best, significantly better than LLaMA and Ernie Bot ($ps < 0.01$), followed by ChatGPT, which was significantly better than LLaMA ($p = 0.01$); no differences existed among other models.

4. Discussion

This study used a reading-based vocabulary learning task (zero-shot learning for LLMs) to compare human learners and LLMs in affective learning during language learning (incidental vocabulary acquisition), addressing three questions: whether LLMs can learn contextual emotion like humans, and how context variability and contextual emotion affect incidental vocabulary acquisition for both LLMs and humans. Results are discussed below in terms of contextual emotion acquisition and word form/meaning acquisition.

4.1 Contextual Emotion Acquisition

Results show that all LLMs and language learners (both L1 and L2) incidentally acquired the contextual emotion of target words through reading exposure: when they encountered target words in positive contexts during reading, they judged those words as emotionally positive in testing, and vice versa for negative contexts. Social psychologists term this behavioral convergence "emotional contagion" —the sharing of emotional states between individuals (Hatfield et al., 1993; Ho et al., 2023)—while psycholinguists call it "transfer," where emotion migrates from context to the target word itself (see Snefjella et al., 2020). The question is: how does this contagion or transfer occur?

To address this, we propose a **Dual-Mechanism Framework**, arguing that humans and LLMs achieve functionally equivalent behavioral patterns through fundamentally different pathways. For human learners, affective learning is an embodied, situated process whose internal mechanisms are powerfully explained by Embodied Cognition Theory. This theory posits that language meaning construction is rooted in individuals'sensorimotor interactions with the environment and bodily experiences (Barsalou, 2008), and heavily depends on embodied simulation triggered by linguistic input. For example, when learners read "Her wonderful teacher performed at the exciting concert," they do not passively receive symbolic text but actively activate perception-action systems and episodic memories associated with positive emotion, imaginatively reconstructing multimodal representations such as excited emotional experiences and warm social interactions. Target words (e.g., *aunith*) become "infected" with emotional coloring through repeated co-occurrence with such rich embodied experiences, achieving contextual emotion transfer. This process is deeply embedded in human biology and social interaction, showing that affective learning is not an independent cognitive module but a dynamic system closely linked to motivational regulation, attention allocation, and deep memory encoding (see Eysenck & Brysbaert, 2018; Pessoa, 2008).

For LLMs, however, affective learning is not dependent on embodied cognitive systems but is a computational process based on statistical distributions and vector space optimization. Unlike humans who construct semantics through sensory and bodily experiences, LLMs' "learning" stems from extracting and reproducing lexical co-occurrence patterns in massive corpora (Bisk et al., 2020; Landauer & Dumais, 1997). This core operation can be formalized as: when processing large amounts of context containing emotional cues, the model continuously adjusts the target word' s vector position in high-dimensional semantic space through self-attention mechanisms. When a novel word (e.g., *aunith*) repeatedly appears in positive contexts, its vector gradually moves closer to semantic clusters of positive words like *exciting* and *wonderful*, thereby mathematically acquiring positive emotional values. Consequently, emotional valence ratings can be interpreted as similarity calculations between target word vectors and emotional anchor vectors, where score magnitude reflects semantic space proximity rather than subjective experience. In sentence generation, word vectors positioned in "positive regions" are more likely to trigger connections to positive vocabulary during autoregressive sampling, systematically reproducing emotional consistency. This mechanism shows that LLMs' emotional transfer is essentially pattern completion based on statistical co-occurrence, not equivalent to human embodied experience. However, this study found that they exhibit "positivity advantage" and other behavioral effects highly consistent with humans. This phenomenon highlights a "heterogeneous isomorphism" mechanism: through completely different technical pathways, purely computational processes relying on distributed semantic modeling and gradient optimization enable LLMs to demonstrate functionally similar affective transfer capabilities. This **Dual-Mechanism Framework** not only provides a new theoretical anchor for understanding AI' s human-like characteristics but also, through computational modeling, reflects the indispensability of embodied experience in human affective learning while reminding us that behavioral similarity cannot be simply equated with mechanistic identity (Binz & Schulz, 2023; Hagendorff et al., 2023).

Notably, from the perspective of Usage-Based Learning Theory (Ellis & Wulff, 2015; Tomasello, 2005), both humans and LLMs share common features of frequency-driven learning and distributional generalization in emotional context learning: both gradually abstract and form stable emotion-semantics connections through repeated exposure to context-word co-occurrence patterns. However, mechanistic differences lurk beneath this surface similarity. For human learners, frequency effects are realized through embodied simulation and social interaction, where frequency strengthens deep encoding between emotional experiences and lexical forms. For LLMs, frequency is directly manifested as statistical distribution optimization of word vectors in high-dimensional space—a computational process based on distributed representations and self-attention mechanisms.

We further found that when LLMs used "learned" target words for sentence production, they not only maintained grammatical and semantic correctness but also reproduced emotional orientations highly consistent with the learning context. This result is particularly crucial: it shows that LLMs have not only established "emotion-word" connections at the representation level but can also reproduce these correspondences during language production, demonstrating "emotional consistency" and "positivity advantage" highly convergent with humans. Theoretically, this phenomenon has "surprising" implications: existing embodied cognition and emotional transfer theories assume sensory experience, bodily engagement, and emotional regulation as prerequisites, never hypothesizing that purely computational systems lacking embodied channels could exhibit analogous human emotional transfer features (Barsalou, 2008; Snefjella et al., 2020). However, this study' s evidence shows that even under zero-shot conditions, LLMs can rely on distributed representations and self-attention mechanisms to conditionally model statistical co-occurrence relationships between context and vocabulary in high-dimensional semantic space, and through probabilistic pattern completion in autoregressive generation, guide target words to matching emotional regions. In other words, LLMs' "human-like" emotional production does not stem from embodied experience but is a product of computational processes. We argue that LLMs' performance not only extends the boundary of "human-like" affective learning from input and representation to output and generation but also provides new testable pathways for theoretical construction of emotion-language interaction. It further validates our proposed **Dual-Mechanism Framework**: human emotional transfer depends on perception-emotion systems driven by embodied experience, while LLMs' similar performance relies on structured modeling and generalized reproduction of linguistic statistical structures.

4.2 Word Form and Semantic Acquisition

The above results show that LLMs and human learners exhibited highly consistent patterns in vocabulary acquisition (word form and meaning), both significantly influenced by contextual emotion and context variability. First, both meaning recognition and generation demonstrated significant positivity advantages: the more positive LLMs and learners judged a target word' s emotion, the better their meaning acquisition performance. Since this positivity originated from contextual emotion, this actually demonstrates that positive contextual emotion promoted meaning acquisition. Psychology has proposed many explanations for positive emotion's facilitative effect on human learning. For example, some scholars suggest positive emotion can motivate learners and enhance attention, thereby promoting sustained interaction with learning materials (MacIntyre & Vincze, 2017). Others propose that positive contexts can serve as mnemonic cues, facilitating retrieval of word meanings from memory. When target words become associated with positive contextual emotion, more vivid and retrievable memory traces are formed (Lana & Kuperman, 2024; Snefjella et al., 2020). LLMs' demonstrated positivity advantage, from the perspective of zero-shot vocabulary learning, reveals LLMs' human-like characteristics while also supporting the view that emotion influences human learning. In this study, the positivity advantage stems from the tight association between emotion and context, which effectively promotes deep processing and precise acquisition of vocabulary.

Second, both meaning recognition and generation demonstrated significant varied context advantages: meaning learning was always better under varied contexts than repeated contexts. This result aligns with the context variability hypothesis and supports the superiority of diverse contexts from an AI perspective. Johns and Jones' s (2008) Semantic Distinctiveness Model (SDM) provides a powerful explanation for this pattern from the perspective of word meaning representation. In SDM, words are part of a word-context matrix' s document distribution. This matrix represents words as vectors using indices such as frequency, document count, or semantic distinctiveness. As words appear in more different contexts, their vector representations expand. During reading, readers continuously compare information stored in vectors with fresh contexts where words appear. If fresh contexts align with already-stored information, they are encoded only weakly. Conversely, if fresh contexts contain new independent information not stored in vector representations, they are encoded more strongly. Each time learners encounter a word in text, they initiate a new cognitive process, comparing the novel word with context knowledge stored in their mental lexicon. If it differs from stored context knowledge, the word' s mental representation is updated; if it aligns with context knowledge stored in memory vectors, the mental representation is updated less. Based on this, the richer the contextual information readers can obtain about a word, the richer their mental lexicon updates (Jones et al., 2017). This helps readers and LLMs acquire decontextualized knowledge of the word, promoting learning of core meaning.

Additionally, in definition generation performance, both LLMs and human learners showed results significantly different from meaning recognition, mainly manifested in the amplification of varied context advantage caused by the significant interaction between context variability and contextual emotion: under varied contexts, as emotional valence increased, both learners' and LLMs' performance improved, widening the gap with repeated contexts. For human learners, this result indicates that varied contexts provide richer situational cues for the same target word, triggering more frequent embodied simulations related to motivation, attention, and deep encoding, while positive contexts further enhance mood-extraction cue matching, forming more stable and retrievable high-level semantic representations in recall-based production (Schmidt, 1990; Tulving &

Thomson, 1973). For LLMs, repeated appearances of the same target word in diverse contexts allow its context-dependent vector representation to acquire more dispersed emotional/semantic signals in high-dimensional semantic space: self-attention aligns different emotional cues across contexts (e.g., co-occurrence with *exciting*, *wonderful* word clusters), thereby pushing the target word vector closer to positive vocabulary clusters and enhancing the linear separability and conditional likelihood of emotional dimensions (see Andrews et al., 2009; Landauer & Dumais, 1997). Additionally, positive contexts have higher lexical density and co-occurrence connectivity in general corpora (positive words are more "dense" in the network), making target words more likely to "fall into" the "attractor basin" of positive regions during training/inference, preferentially sampling descriptive language consistent with that region in autoregressive generation, thus demonstrating stronger "emotional consistency" and higher accuracy in demanding recall-based outputs like definition generation (see Clark et al., 2019; Radford et al., 2019). In other words, humans' amplification effect stems from deep encoding and retrieval matching driven by embodied simulation, while LLMs' amplification effect originates from conditional aggregation of emotional co-occurrence signals and high-density neighborhood effects in vector geometry under varied contexts.

It should be noted, however, that although LLMs and human learners showed high similarity in overall patterns of vocabulary acquisition (word form and meaning), LLMs demonstrated significantly superior accuracy on specific test measures, including word form tests, definition matching, and definition generation. We argue that this advantage does not indicate LLMs are "smarter" than humans but reveals fundamental mechanistic differences. Human learning is a resource-limited process strongly influenced by working memory capacity, attention fluctuations, motivation levels, and existing knowledge structures (Eysenck & Brysbaert, 2018). In limited experimental exposure, human learners may not fully integrate all information. In contrast, LLMs, as computational architectures centered on prediction and pattern matching, nearly perfectly execute statistical optimization tasks within their context windows. They are unaffected by fatigue, distraction, or motivation issues and can unwaveringly utilize every statistical cue in input text (Bisk et al., 2020). Specifically, under the current task framework, LLMs behave like idealized statistical learners without cognitive constraints. Their stronger, purer "context variability advantage" and "positivity advantage" can be viewed as powerful validation of the distributional semantics hypothesis—that lexical meaning arises from contextual distribution. LLMs amplify these effects, precisely proving these learning principles are efficient and powerful. Therefore, LLM performance uniquely demonstrates that similar effects observed in human vocabulary acquisition likely follow equally efficient statistical learning principles at their foundation, merely constrained by human biological and cognitive conditions. This "machine surpassing human" finding strengthens our proposed **Dual-Mechanism Framework** and highlights the potential of using LLMs as computational models to isolate and test variables in human learning models.

It should be noted that although this study included four representative LLMs for comparative analysis, results still have certain generalizability limitations. First, the four models differ substantially in architecture and scale, encompassing standard Transformer architectures (e.g., LLaMA and ChatGPT) and hybrid architectures integrating MoE mechanisms (e.g., Gemini), whose differences in reasoning paths and parameter activation mechanisms may affect contextual emotion processing. Second, language optimization directions differ: Ernie Bot is Chinese-optimized, while ChatGPT and Gemini are primarily English-trained, which may cause language-specific differences in models' responses to emotional signals in contexts. Therefore, although we achieved uniformity in input structure and test task design, differences in language adaptability and semantic representation mechanisms across models require further validation of consistency and robustness in contextual emotion acquisition abilities in future research. Additionally, the current study only selected four representative LLMs; future research should expand model coverage to include more architectural types and corpus backgrounds to systematically examine stability and variability across models in contextual emotion acquisition tasks.

Conclusion

This study employed a zero-shot language learning framework to compare four representative LLMs and three human learner groups in affective learning performance during reading-based incidental vocabulary acquisition, as well as the effects of contextual emotion and variability on learning outcomes. Results showed that LLMs can not only "absorb" and transfer emotion from contexts like humans but also exhibit "positivity advantage" and "context variability advantage" highly consistent with humans in word form and meaning learning, even surpassing human learners in accuracy on multiple measures. These findings indicate that LLMs possess human-level contextual sensitivity and emotional consistency, highlighting the power of distributed semantic modeling in complex learning tasks.

At the theoretical level, we proposed and validated a **Dual-Mechanism Framework** to explain the heterogeneous cognitive foundations behind behavioral similarity: human affective learning depends on perception-emotion systems driven by embodied experience, deeply embedded in sensorimotor experiences, episodic memory, and social interaction; LLMs' emotional transfer, however, originates from statistical modeling and reproduction of linguistic co-occurrence patterns in high-dimensional semantic space—a computational process based on self-attention and vector space optimization. Both yield functionally similar results but reflect fundamental differences between biological cognition and machine computation. This framework provides theoretical tools for understanding "human-like" emotional performance in AI and offers a systematic comparative foundation for exploring cross-mechanism implementation paths for language-emotion interaction.

More importantly, this study reveals at the macro level the double-edged significance of affective learning in AI development. On one hand, LLMs' human-like affective learning and transfer capabilities bring unprecedented opportunities to human-AI interaction in education, healthcare, and social governance, as they can capture and reproduce context-word emotional connections without affective annotation, thereby enhancing naturalness and acceptability of generated content. On the other hand, this capability also alerts us to ethical risks: when models inherit and amplify cultural biases and affective tendencies from corpora, they may unconsciously influence the value orientation of their emotional outputs (Caliskan et al., 2017). In other words, LLMs'"emotional consistency" is not only experimental validation for cognitive science and distributional semantics theories but also a frontier issue in AI ethics assessment.

In summary, this study' s findings not only advance our deep understanding of human and machine mechanisms in affective learning but also provide solid empirical evidence and theoretical anchors for future development of emotional norms and risk prevention strategies for AI systems.

References

Ahmed, S. (2004). *The cultural politics of emotion*. New York: Routledge.

Ahmed, S. (2010). *The promise of happiness*. London: Duke University Press.

Andrews, B., Vigliocco, G., & Vinson, D. P. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review, 116*(3), 463−498.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390−412.

Balass, M. (2011). *Learning words in context: An ERP investigation of word experience effects on familiarity and meaning acquisition* (Unpublished doctoral dissertation). University of Pittsburgh.

Barrett, L. F. (2017). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience, 12*(11), 1833.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59*(1), 617−645.

Binz, M., & Schulz, E. (2023). Turning large language models into cognitive models. *Computer Science*. Advance online publication. https://doi.org/10.48550/arXiv.2306.03917

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., ··· Turian, J. (2020). Experience grounds language. *Computer Science*. Advance online publication. https://doi.org/10.48550/arXiv.2004.10151

Blythe, H. I., Liang, F., Zang, C., Wang, J., Yan, G., Bai, X., & Liversedge, S. P. (2012). Inserting spaces into Chinese text helps readers to learn new words:

An eye movement study. *Journal of Memory and Language, 67*(2), 241—254.

Bolger, D. J., Balass, M., Landen, E., & Perfetti, C. A. (2008). Context variation and definitions in learning the meanings of words: An instance-based learning approach. *Discourse Processes, 45*(2), 122—159.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ··· Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877—1901.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183—186.

Chomsky, N. (1957). Fundamentals of language. *International Journal of American Linguistics, 23*(3), 234—242.

Christensen, R. H. B. (2023). *ordinal: Regression models for ordinal data.* R package version 2023.12-4.1. https://CRAN.R-project.org/package=ordinal

Clark, E., Celikyilmaz, A., & Smith, N. A. (2019, July). Sentence mover' s similarity: Automatic evaluation for multi-sentence texts. Paper presented at the meeting of the Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason and the human brain.* New York: Grosset/Putnam.

Driver, M. (2022). Emotion-laden texts and words: The influence of emotion on vocabulary learning for heritage and foreign language learners. *Studies in Second Language Acquisition, 44*(4), 1071—1094.

Ellis, N. C., & Wulff, S. (2015). Usage-based approaches to SLA. In B. VanPatten, & J. Williams (Eds.), *Second language acquisition research series: Theories in second language acquisition* (pp. 75—94). Routledge.

Eysenck, M. W., & Brysbaert, M. (2018). *Fundamentals of cognition* (3rd ed.). Routledge.

Godfroid, A., Ahn, J., Choi, I., Ballard, L., Cui, Y., Johnston, S., ···Yoon, H. -J. (2018). Incidental vocabulary learning in a natural reading context: An eye-tracking study. *Bilingualism: Language and Cognition, 21*(3), 563—584.

Hagendorff, T., Dasgupta, I., Binz, M., Chan, S. C. Y., Lampinen, A., Wang, J. X., ···Schulz, E. (2023). Machine psychology. *Computer Science.* Advance online publication. https://doi.org/10.48550/arXiv.2303.13988

Hasson, U., Ghazanfar, A.A., Galantucci, B., Garrod, S., & Keysers, C. (2012). Brain-to-brain coupling: A mechanism for creating and sharing a social world. *Trends in cognitive sciences, 16*(2), 114—121.

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1993). *Emotional Contagion* (1st ed.). New York: Cambridge University Press.

Ho, M. H., Kemp, B. T., Eisenbarth, H., & Rijnders, R. J. P. (2023). Designing a neuroclinical assessment of empathy deficits in psychopathy based on the Zipper Model of Empathy. *Neuroscience and Biobehavioral Reviews, 151*, 105244.

Horst, J. S., Parsons, K. L., & Bryan, N. M. (2011). Get the story straight: Contextual repetition promotes word learning from storybooks. *Frontiers in Psychology, 2*, 17.

Hulstijn, J. H. (2001). Intentional and incidental second-language vocabulary learning: A reappraisal of elaboration, rehearsal and automaticity. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 258−286). Cambridge University Press.

Johns, B. T., & Jones, M. N. (2008). Predicting lexical decision and naming times from a semantic space model. *Proceedings of the 30th Annual Cognitive Science Society, 30*(30), 279−284.

Jones, M. N., Dye, M., & Johns, B. T. (2017). Context as an organizing principle of the lexicon. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 67, pp. 239−283). United States: Elsevier Science & Technology.

Joseph, H., & Nation, K. (2018). Examining incidental word learning during reading in children: The role of context. *Journal of Experimental Child Psychology, 166*, 190−211.

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods, 42*(3), 627−633.

Lana, N., & Kuperman, V. (2024). Learning concrete and abstract novel words in emotional contexts: Evidence from incidental vocabulary learning. *Language Learning and Development, 20*(2), 158−173.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato' s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211−240.

Laufer, B., & Aviad-Levitzky, T. (2017). What type of vocabulary knowledge predicts reading comprehension: Word meaning recall or word meaning recognition? *The Modern Language Journal, 101*(4), 729−741.

Lauro, J., Schwartz, A. I., & Francis, W. S. (2020). Bilingual novel word learning in sentence contexts: Effects of semantic and language variation. *Journal of Memory and Language, 113*, 104123.

Li, Z. (2024). *Semantic prosody acquisition and its influence on the learning of L2 novel word forms and meanings* (Unpublished doctoral dissertation). Shanghai Jiao Tong University. [李赞. (2024). 语义韵习得及其对二语新颖词词形和词义学习的影响 (博士学位论文). 上海交通大学.]

Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis, & E. Tognini-Bonelli

(Eds.), *Text and technology: In honour of John Sinclair* (pp. 157–176). Amsterdam, The Netherlands: John Benjamins.

Ma, Z., & Li, Z. (2024). Acquiring semantic prosody in L2 novel word learning: The effect of context variability and gender. *Modern Foreign Languages, 47*(6), 790–801. [马拯, 李赞. (2024). 二语新颖词语义韵的习得：语境变异性及性别的影响. 现代外语, 47(6), 790–801.]

MacIntyre, P. D., & Vincze, L. (2017). Positive and negative emotions underlie motivation for L2 learning. *Studies in Second Language Learning and Teaching, 7*(1), 61–88.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81–97.

Nevisi, R. B., Hosseinpur, R. M., & Darvish, F. Z. (2018). The impact of L1/L2-based explicit output task instruction on Iranian EFL learners' semantic prosody learning. *Journal of Language Horizons, 2*(2), 51–74.

Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature Reviews Neuroscience, 9*(2), 148–158.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1*(8), 9.

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*(2), 129.

Sinclair, J. M. (1987). *Looking up.* London: Collins.

Snefjella, B., Lana, N., & Kuperman, V. (2020). How emotion is learned: Semantic learning of novel words in emotional contexts. *Journal of Memory and Language, 115*, 104171.

Stewart, J., Gyllstad, H., Nicklin, C., & McLean, S. (2024). Establishing meaning recall and meaning recognition vocabulary knowledge as distinct psychometric constructs in relation to reading proficiency. *Language Testing, 41*(1), 89–108.

Tamir, M., Schwartz, S. H., Cieciuch, J., Riediger, M., Torres, C., Scollon, C., ···Vishkin, A. (2016). Desired emotions across cultures: A value-based account. *Journal of personality and social psychology, 111*(1), 67–82.

Tomasello, M. (2005). *Constructing a language: A usage-based theory of language acquisition* (1st ed.). Cambridge: Harvard University Press.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80*(5), 352–373.

Wang, W., Zheng, V. W., Yu, H., & Miao, C. (2019). A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology, 10*(2), 1–37.

Wetherell, M. (2012). *Affect and emotion: A new social science understanding* (1st ed.). Los Angeles: SAGE.

Wu, S. Y., & Li, Z. (2024). How semantic prosody is acquired in novel word learning: Evidence from the "Double-Jujube Tree" effect. *Acta Psychologica Sinica, 56*(5), 531−543. [吴诗玉, 李赞. (2024). 新颖词语义韵的发生机制:"双枣树"效应的证据. 心理学报, 56(5), 531−543.]