Do large language models resolve semantic ambiguities in the same way as humans? The case of word segmentation in Chinese sentence reading

Weiyan Liao
(u3007255@connect.hku.hk)
Department of Psychology,
University of Hong Kong

Zixuan Wang
(zwang524-c@my.cityu.edu.hk)
Department of Computer Science,
City University of Hong Kong

Kathy Shum
(kkmshum@hku.hk)
Department of Psychology,
University of Hong Kong

Antoni B. Chan
(abchan@cityu.edu.hk)
Department of Computer Science,
City University of Hong Kong

Janet H. Hsiao
(jhhsiao@ust.hk)
Division of Social Science,
Hong Kong University of Science and Technology

Abstract

Large language models (LLMs) were trained to predict words without having explicit semantic word representations as humans do. Here we compared LLMs and humans in resolving semantic ambiguities at the word/token level by examining the case of segmenting overlapping ambiguous strings in Chinese sentence reading, where three characters "ABC" could be segmented in either "AB/C" or "A/BC" depending on the context. We showed that although LLMs performed worse than humans, they demonstrated a similar interaction effect between segmentation structure and word frequency order, suggesting that this effect observed in humans could be accounted for by statistical learning of word/token occurrence regularities without assuming an explicit semantic word representation. Nevertheless, across stimuli LLMs' responses were not correlated with any human performance or eye movement measures, suggesting differences in the underlying processing mechanisms. Thus, it is essential to understand these differences through XAI methods to facilitate LLM adoption.

Keywords: EMHMM; eye tracking; large language models; reading; Chinese

Introduction

Large language models (LLMs) are trained to predict words from massive datasets of text. Although their predictions appear to follow the semantic meaning of the inputs in most cases and exhibit human-like behavior in sentence understanding, LLMs did not learn about the semantic representation of words explicitly as humans do. More specifically, humans semantic representations of words have been shown to be highly associated with sensory-motor experiences (e.g., Bedny & Caramazza, 2011). In contrast, LLMs use word/token co-occurrence regularities to process sentences, which lack sensory-motor representations of the words. Wang et al. (2020) found that during object color recognition tasks, both sighted and blind participants showed activation in the left dorsal anterior temporal lobe, whereas only sighted participants showed activation in the ventral occipitotemporal color perceptual region. This finding suggested there are at least two forms of word representations in humans: sensory-derived and language- and cognition-derived. In addition, although blind participants had similar performance to sighted participants, they had greater individual variability, perhaps due to individual differences in compensatory strategies. Similarly, without any explicit sensory representation of words, LLMs' sentence processing and linguistic decisions may depend mainly on statistical regularities of word/token co-occurrences learned from large datasets, and thus may differ from humans' regardless of their human-like behavior in sentence understanding. In addition, humans process words in a sentence sequentially due to the constraints from speech production or limitations of human visual attention in the case of reading (Rayner et al., 2013), whereas LLMs can process all words in a sentence context simultaneously when learning to predict the next word in a sentence. This difference may also lead to differences in how words are represented and processed in humans vs. LLMs. Here we aimed to investigate this possibility through examining a special case of word- or token-level semantic ambiguity resolution, word segmentation in Chinese sentence reading.

Semantically ambiguous sentences are common in natural languages. However, humans do not typically have difficulties in resolving these ambiguities in daily life. To examine whether LLMs have similar semantic disambiguation abilities as humans, Liu et al. (2023) presented LLMs with ambiguous English sentences with more than one interpretations (for instance, the sentence "The cat was lost after leaving the house" can be interpreted as "the cat was unable to find its own way" or "the cat was unable to be found"). They asked LLMs to generate disambiguations for a sentence, and then recruited humans to evaluate LLMs' performance by majority voting. They found that the disambiguations generated by GPT-4, which performed the best among the LLMs in their study, were considered correct only 32% of the time in human evaluation, much lower than human performance in this task. This result suggested that sentence disambiguation may be a difficult task for LLMs as compared with humans.

Instead of sentence-level semantic disambiguation, where a word may have multiple meanings and the resolution depends on the sentence context, here we focused on word- or token-level semantic disambiguation. Word segmentation in Chinese sentence reading provided a unique opportunity for this examination. More specifically, in Chinese sentence reading, there is no word boundary information such as space used in other languages. Thus, full understanding of Chinese sentences requires correct segmentation at the word/token level. In a sentence with a three-character overlapping am-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

biguous string (OAS), denoted as "ABC", in which the middle character could form distinct words with the characters on both its left (word AB) and its right (word BC), different segmentations could lead to different interpretations of the sentence, and typically only one of them would be consistent with the semantics of the sentence. For instance, the OAS 表明早 could be parsed as 表明/早 (*show* and *early*) or 表/明早 (*form* and *tomorrow morning*). However, in the sentence in Figure 1a, only the segmentation on the top was correct, and the sentence was interpreted as "the experiment results show that discovering and treating the disease early can prevent a major disease". In previous studies (Huang & Li, 2020; Huang et al., 2021), the average comprehension accuracy was 95% in human participants (Chinese readers), showing that humans could perform well in this word-level disambiguation task. Since LLMs do not have explicit word semantic representations and their word-level semantic processing depends mainly on word/token co-occurrence regularities learned from large datasets, we expected that they may make more mistakes than humans in this task. We also speculated that they may process these OASs differently from humans, which could lead to different word segmentation decisions across OAS conditions or individual trials.

Eye tracking has been commonly used to examine cognitive processes involved in sentence reading. In disambiguating OASs, Huang and Li (2020) reported that participants typically had shorter first-fixation duration, shorter gaze duration, and shorter second-pass reading time in the OAS region and shorter total sentence reading time in the AB/C than in the A/BC segmentation structure (Huang & Li, 2020). This result suggested that humans' cognitive processing was influenced by segmentation structure. In addition, the word frequency of word AB and BC in an OAS appeared to influence readers' processing efficiency, as participants had shorter fixation duration and shorter first-pass and go-pass reading time when words in the OAS region followed the high-low frequency order (when word AB had higher frequency than BC) than the low-high order (Huang et al., 2021). These effects may be related to humans' left-to-right sentence reading direction.

Accordingly, in the current study, we measured human readers' cognitive processes involved in disambiguating OASs using gaze duration and first-fixation duration in the OAS region. In addition, we used Eye Movement analysis with Hidden Markov Models (EMHMM; Chuk et al., 2014) to quantify human readers' word segmentation processes. More specifically, EMHMM is a machine learning approach for eye movement analysis. By assuming two models with different word segmentation structures (Figure 1), we may use the log-likelihoods of a participant's eye movement data being generated by the models to quantify how well one's eye movement behavior follows a particular word segmentation interpretation. To compare with LLMs' processing of OASs, we measured LLMs' preference for the correct segmentation structure for sentences with an OAS, as measured via sentence likelihoods. We then examined the effect of segmentation structure and frequency order on participants' word segmentation performance and eye movement behavior, and compared them with LLMs' accuracy and preference for the correct word segmentation structure. We hypothesized that humans' performance and eye movement behavior would be influenced by frequency order and segmentation structure in the OAS region as in previous studies, and there may be an interaction between frequency order and segmentation structure where the advantage of having a high-low frequency order may be larger when the correct segmentation structure is AB/C. In contrast, since LLMs may have a better ability to process multiple words simultaneously, they may exhibit a weaker segmentation structure or a frequency order main effect. However, they may still exhibit an interaction effect between frequency order and segmentation structure, where the preference for the correct segmentation structure would be higher when the correct segmentation structure matched the word frequency order (i.e., AB/C in the high-low frequency order condition, and A/BC in the low-high frequency order condition), and this word frequency effect could be learned purely from statistics of word occurrences in the training dataset (e.g., Schepens et al., 2023). Also, since LLMs do not have explicit sensory-motor representation of the words, they may not perform as well as humans in word segmentation.

	1	2	3	4	5	6	7	8	9	10
Prior	.13	.12	.08	.09	.15	.10	.08	.08	.11	.05
1	.10	.11	.11	.10	.11	.10	.09	.09	.10	.09
2	.11	.10	.11	.10	.10	.10	.10	.09	.09	.09
3	.10	.12	.10	.10	.11	.10	.10	.09	.10	.09
4	.11	.11	.11	.10	.10	.10	.10	.09	.09	.09
5	.10	.09	.13	.09	.10	.09	.11	.10	.10	.09
6	.09	.10	.11	.10	.11	.09	.10	.09	.11	.09
7	.10	.12	.09	.09	.12	.10	.10	.09	.09	.09
8	.10	.11	.11	.11	.10	.10	.10	.10	.10	.10
9	.10	.10	.10	.10	.10	.10	.11	.10	.10	.10
10	.10	.10	.10	.10	.10	.10	.10	.10	.10	.10

Figure 1: (a) Examples of predefined ROIs in the AB/C pattern (top) and the A/BC pattern (bottom). (b) Example HMM summarizing a participant's gaze transition pattern during reading. Ellipses show predefined ROIs as 2-D Gaussian emissions and represent 2 SD from the mean. The color of the fixation shows the assignment to its ROI. Table on the bottom shows transition probabilities among the ROIs. Priors show the probabilities that a fixation sequence starts from the ellipse.

Methods

Participants

We recruited 65 adult native Chinese readers as the participants (46 females). According to a power analysis, a sample size of 52 was needed to acquire a large effect size (following previous studies on Chinese OAS, e.g., Huang & Li, 2020; Huang et al., 2021) in a 2 x 2 repeated ANOVA (f = .40; β = .2; a = .05). A sample size of 55 was needed to acquire a large effect size in a linear multiple regression with one tested

predictor (f^2 = .35; β = .2; a = .05). The participants were aged from 18-30 (M = 21.508, SD = 2.873).

Materials

The Chinese sentence reading task consisted of 80 sentences. We first chose from Huang et al. (2021) 40 3-character traditional Chinese OASs, each of which could be parsed as either AB/C or A/BC segmentation structure depending on the context. Word frequency information was from the Text Corpus for Word frequency in Contemporary Chinese (Institute of Linguistics, Academia Sinica, 2005). In 20 of the OASs, the word AB had higher word frequency (M = 141.48 occurrences per million) than the word BC (M = 1.58 occurrences per million), $t(35) = 12.30$, $p < .001$. We referred to these as OASs with a high-low frequency order. In the other 20 OASs, the word AB had lower word frequency (M = 1.44 occurrences per million) than the word BC (M = 167.39 occurrences per million), $t(35) = -8.62$, $p < .001$. These OASs were referred to as having a low-high frequency order. Each OAS appeared in two Chinese sentences: in one sentence the correct segmentation of the OAS was AB/C, whereas in the other the correct segmentation was A/BC. Each sentence contained 10 to 20 characters (M = 15.201). Sentence lengths and average character complexity defined by number of strokes in a sentence were matched across the conditions. To perform by-item analysis, according to a power analysis, a sample size of 32 items was needed to acquire a large effect size in a 3 x 2 x 2 between-within ANOVA (f = .40, β = .2; a = .05). A sample size of 56 was needed to acquire a large effect size in a 2 x 2 between-subject ANOVA (f = .40, β = .2; a = .05).

Design

In human data analysis, the design consisted of two within-participant variables, frequency order (high-low vs. low-high) and segmentation structure (AB/C vs. A/BC). The dependent variables were performance as measured in accuracy and reaction time (RT) when judging the segmentation structure through key responses, and eye movement measures during natural sentence reading (the first sentence presentation in Figure 2; see Procedure for details) including OAS total gaze duration (the sum of the duration of all fixations within the OAS region), OAS first-fixation duration (the duration of the first-fixation landing into the OAS region), and eye movement pattern as quantified using EMHMM (see EMHMM section for details). Linear regression analyses were used to examine whether eye movement measures could predict reading performance.

To examine whether LLMs exhibited similar behavior as humans, we measured the preference of the LLMs for the correct segmentation structure, i.e., AB/C or A/BC, by comparing the LLM's log-likelihoods of the two possible segmentations. Specifically, given the same stimuli as the human study, we inserted the symbol "/" to indicate two possible segmentation structures of the OAS.[1] The LLM's log-likelihoods of these two sentences were calculated, and the log-likelihoods were converted into posterior probabilities indicating the LLMs preference for the correct segmentation. We then compared Davinci-003 (Kalyan, 2024) and GPT-3.5 Turbo Instruct (GPT-3.5-TI; Singh, 2023) and examined the effects of segmentation structure and frequency order through a by-item ANOVA with model as a between-subject factor and segmentation structure and frequency order as within-subject factors on the preference for correct segmentation structure.

To compare LLM and human behavior, we directly compared LLMs' preference for the correct segmentation structure with human accuracy in selecting the correct segmentation structure. A 3 x 2 x 2 between-within ANOVA was used to examine the effect of group (human vs. Davinci-003 vs. GPT-3.5-TI), segmentation structure, and frequency order on the preference/accuracy. Linear regression was used to examine whether LLMs' preference of correct segmentation structure could predict humans' segmentation performance and eye movement behavior.

Procedure

In the Chinese sentence reading task, participants started with a practice trial, followed by four experimental blocks, with each block containing 20 sentences. The experimenter ensured that participants understood the task and answered the practice trial correctly before they proceeded to the experimental blocks. At the beginning of each block, a nine-point calibration procedure was performed. Each trial started with a solid dot at the screen center for drift check, followed by a cross on the left side of the screen. Participants were instructed to look at the cross when it appeared. After a fixation was detected at the cross, the cross disappeared, and the target sentence was presented 1° of visual angle to the right of the cross. Participants read the sentence naturally and pressed the spacebar to indicate finishing reading. The cross then appeared again on the left side of the screen and participants were instructed to look at the cross again. After a fixation was detected at the cross, the cross disappeared, and the same sentence with the OAS highlighted in red was presented 1° of visual angle to the right of the cross. Participants pressed the key "F" if they thought the correct segmentation structure of the OAS was AB/C, otherwise pressed the key "J" (Figure 2).

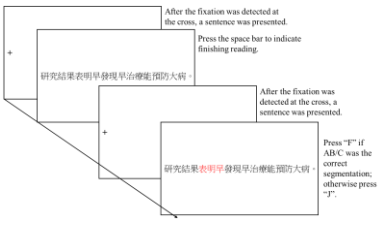

Figure 2: Procedure of the Chinese sentence reading task.

[1] Using "/" led to the largest difference in the posterior probability between the two options among a series of symbols, numbers, English letters, and Chinese characters that we tested.

EMHMM

We used EMHMM with fixed-ROIs to quantify participants' segmentation processes as reflected in eye movement behavior (e.g., Cho et al., 2022a; 2022b; 2022c; Chuk et al., 2020; Lo et al., 2023). Following previous eye movement studies on reading where ROIs were typically pre-defined to be on individual words with the assumption that words are the basic functional units of sentence processing (e.g., Perfetti, 1985), we predefined an ROI on each word except for the OAS region in a sentence. In order to quantify participants' eye movement patterns along the contrast between the two segmentation structure interpretations AB/C and A/BC, we constructed two models, with the ROIs in the OAS regions following the two segmentation structures respectively (Figure 1a). Then, we summarized each participant's gaze transitions during reading each sentence using an HMM in terms of the predefined ROIs and transition probabilities among the ROIs (Figure 1b), for each of the two segmentation structures respectively. In the HMM, each hidden state corresponded to an ROI, and fixations within an ROI were assumed to follow a Gaussian distribution. The representative models of AB/C and A/BC segmentations were then formed by summarizing all individual AB/C and A/BC models respectively[2]. For each participant, we quantified the eye movement pattern during reading each sentence along the dimension contrasting the AB/C and A/BC segmentation patterns using L-R (Left vs. Right word segmentation) scale, which is defined as (L - R) / (|L| + |R|), where L refers to the log-likelihood of the participant's eye movement data under the HMM of the AB/C pattern, and R refers to the log-likelihood of the participant's data under the HMM of the A/BC pattern (e.g., An & Hsiao, 2021; Chan et al., 2018; Chan, Suen et al., 2020; Chan, Barry et al., 2020; Hsiao, An et al., 2021; Liao et al., 2022; Zhang et al., 2019; Zheng et al., 2022). A more positive L-R scale indicates greater similarity to the AB/C pattern.

Results

Human data

In accuracy, the results revealed a main effect of segmentation structure, $F_1(1, 64) = 18.173$, $p < .001$, $\eta_p^2 = .221$ (it was not significant in the by-item analysis, $F_2(1, 76) = 2.182$, $p = .144$): participants had higher accuracy in the AB/C than the A/BC segmentation structure condition. This effect interacted with frequency order, $F_1(1, 64) = 53.022$, $p < .001$, $\eta_p^2 = .453$, $F_2(1, 76) = 9.175$, $p = .003$, $\eta_p^2 = .108$: in the high-low frequency order condition, participants had higher accuracy in the AB/C than the A/BC segmentation structure, $t_1(64) = 7.665$, $p < .001$, $t_2(76) = 3.186$, $p = .011$; this pattern was reversed in the low-high condition, $t_1(64) = -2.895$, $p = .026$, $t_2(76) = -1.097$, $p = .692$ (Figure 3a).

In RT, a main effect of segmentation structure in the by-subject analysis was observed, $F_1(1, 64) = 11.495$, $p = .001$, $\eta_p^2 = .152$ (it was not significant in the by-item analysis, $F_2(1, 76) = 2.113$, $p = .150$): participants had shorter RT in the AB/C than the A/BC segmentation structure (Figure 3b).

In OAS total gaze duration, the results revealed a main effect of segmentation structure, $F_1(1, 64) = 6.384$, $p = .013$, $\eta_p^2 = .092$, (it was not significant in the by-item analysis, $F_2(1, 76) = 1.512$, $p = .223$): participants had shorter OAS total gaze duration in the AB/C than the A/BC segmentation structure. This effect interacted with frequency order, $F_1(1, 64) = 77.027$, $p < .001$, $\eta_p^2 = .546$, $F_2(1, 76) = 20.641$, $p < .001$, $\eta_p^2 = .214$: in the high-low condition, participants had shorter OAS total gaze duration in the AB/C than the A/BC segmentation structure, $t_1(64) = -8.728$, $p < .001$, $t_2(76) = -4.082$, $p = .011$; this pattern was reversed in the low-high condition, $t_1(64) = 5.291$, $p < .001$, $t_2(76) = 2.343$, $p = .097$ (Figure 3c).

In OAS first-fixation duration, the results revealed a main effect of segmentation structure, $F_1(1, 64) = 7.089$, $p = .010$, $\eta_p^2 = .100$, $F_2(1, 76) = 5.093$, $p = .027$, $\eta_p^2 = .063$: participants had shorter OAS first-fixation duration in the AB/C than the A/BC segmentation structure. This effect interacted with frequency order, $F_1(1, 64) = 8.727$, $p = .004$, $\eta_p^2 = .120$, $F_2(1, 76) = 6.698$, $p = .012$, $\eta_p^2 = .081$: in the high-low frequency order, participants had shorter OAS first-fixation duration in the AB/C than the A/BC segmentation structure, $t_1(64) = -3.955$, $p = .001$, $t_2(76) = -3.314$, $p = .008$; this effect was not observed in the low-high condition $t_1(64) = -.274$, $p = .993$, $t_2(76) = .462$, $p = .967$ (Figure 3d).

In the L-R scale, no effect was observed (Figure 3e).

Linear regression showed that L-R scale was a marginally significant predictor for average RT across trials, $\beta = -.243$, $p = .053$, with $F(1, 62) = 3.895$, $p = .053$, $R^2 = .059$ (Figure 3f).

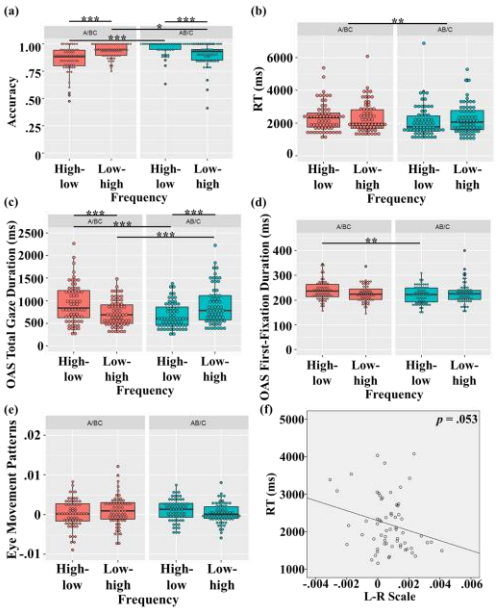

Figure 3: (a) Accuracy, (b) RT, (c) OAS total gaze duration, (d) OAS first-fixation duration, (e) Eye movement pattern as assessed in L-R scale, and (f) Correlation between L-R scales and average RT.

[2] A better method to derive the representative models is to use hidden Markov mixture model (Zhang et al., 2023).

LLM data

In LLMs' preference for the correct segmentation structure, the results revealed an interaction between frequency order and segmentation structure (by-item analysis), $F_2(1, 65) = 5.373$, $p = .024$, $\eta_p^2 = .076$: in the high-low frequency order, LLMs had marginally higher preference for correct segmentation structure in the AB/C than the A/BC segmentation structure, $t_2(65) = 2.434$, $p = .081$. In addition, there was a marginal interaction between frequency order, segmentation structure, and groups, $F_2(1, 65) = 3.016$, $p = .087$, $\eta_p^2 = .044$, whereas there was no main effect of groups, $F_2(1, 65) = 0.182$, $p = .671$. It showed that Davinci-003 and GPT-3.5-TI did not differ significantly in overall performance, but the interaction between frequency order and segmentation structure was stronger in Davinci-003 ($F_2(1, 65) = 9.564$, $p = .003$, $\eta_p^2 = .128$) than GPT-3.5-TI (*n.s.*; Figure 4). There was no main effect of segmentation structure, $F_2(1, 65) = 1.217$, $p = .274$, which was different from human data.

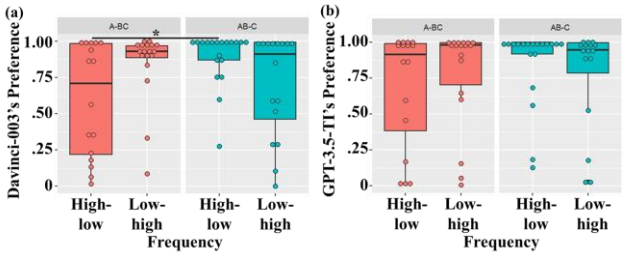

Figure 4: LLMs' preference for correct segmentation structure: (a) Davinci-003, and (b) GPT-3.5-TI.

Comparison between humans and LLMs in performance

LLMs' accuracy was determined according to their calculated preferences. If the preference for the correct segmentation structure was higher than the preference for the wrong segmentation structure for a given sentence, it would be marked as a correct response. The results showed that humans had higher accuracy than LLMs (Table 1).

Table 1: Accuracy mean and SD of humans, Davinci-003, GPT-3.5-TI in each condition. SD are in the parentheses.

	Human	Davinci-003	GPT-3.5-TI
High-low AB/C	.967 (.049)	.947 (.229)	.895 (.315)
High-low A/BC	.860 (.120)	.563 (.512)	.688 (.479)
Low-high AB/C	.900 (.147)	.750 (.447)	.813 (.403)
Low-high A/BC	.937 (.085)	.889 (323)	.833 (.383)

When we compared humans' accuracy with LLMs' preference for correct segmentation structure in the by-item analysis, the results revealed a main effect of group, $F_2(2, 130) = 8.465$, $p < .001$, $\eta_p^2 = .115$: humans' accuracy was higher than Davinci-003's preference, $t_2(65) = 3.742$, $p = .001$, and GPT-3.5-TI's preference, $t_2(65) = 3.105$, $p = .008$. In addition, an interaction between frequency order and segmentation structure was observed, $F_2(1, 65) = 8.334$, $p = .005$, $\eta_p^2 = .114$: in the high-low frequency order, subjects had higher accuracy/preference in the AB/C than the A/BC segmentation structure, $t_2(65) = 3.019$, $p = .019$. There was no interaction among group, frequency order, and segmentation structure, $F_2(2, 130) = 1.854$, $p = .161$, suggesting that humans and LLMs did not significantly differ in this interaction effect between frequency order and segmentation structure. Also, no other main effect or interaction effect was found.

Processing similarities between LLMs and humans

Linear regression was used to test whether LLM's preference for correct segmentation structure could predict human word segmentation processes revealed in eye movement behavior. The results showed that GPT-3.5-TI's preference for correct segmentation structure was a marginally significant predictor for L-R scale, $\beta = .233$, $p = .054$, with $F(1, 67) = 3.845$, $p = .054$, $R^2 = .054$ (Figure 5a), whereas Davinci-003's preference was not a significant predictor for L-R scale, $\beta = .013$, $p = .298$, with $F(1, 67) = 1.100$, $p = .298$, $R^2 = .016$ (Figure 5b). GPT-3.5-TI's and Davinci-003's preference for correct segmentation structure could not predict humans' OAS total gaze duration or OAS first-fixation duration. In addition, they were not a significant predictor for human word segmentation performance in accuracy or RT.

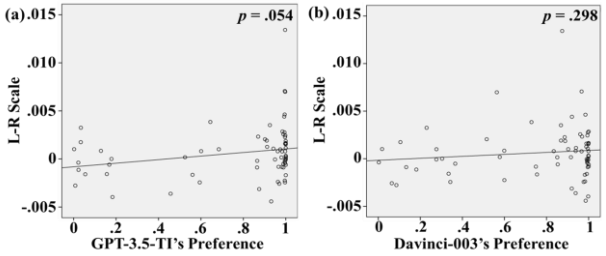

Figure 5: correlation between L-R scale and LLMs' preference for correct segmentation structure: (a) Correlation between L-R and GPT-3.5-TI's preference, and (b) correlation between L-R and Davinci-003's preference.

Discussion

Here we tested whether LLMs resolved semantic ambiguities in the same way as humans in word segmentation during Chinese sentence reading. More specifically, in segmenting words in OASs in a sentence context, previous research has reported that Chinese readers exhibited a segmentation structure effect (i.e., an advantage in processing AB/C over A/BC) and a frequency order effect (i.e., an advantage when AB is of higher word frequency than BC) since humans generally process words in a sentence sequentially from left to right. In contrast, LLMs are able to process multiple words in a sentence simultaneously, and thus may exhibit a weaker effect. In contrast, both humans and LLMs may exhibit an interaction effect between frequency order and segmentation structure (i.e., when the correct segmentation is AB/C, a processing advantage if AB is of higher word frequency than BC, and vice versa if the correct segmentation is A/BC) since this word frequency effect may be learned purely from statistics

of word occurrences in the training dataset. In addition, LLMs may not perform as well as humans in resolving OASs since they did not learn about the semantic representation of words as explicitly as humans do.

Our human data showed a significant main effect of segmentation structure in accuracy, RT, OAS total gaze duration, and OAS first-fixation duration in the by-subject analysis, although it was not significant in the by-item analysis of some measures. Also, a higher similarity of participants' eye fixation behavior to the AB/C segmentation structure interpretation as indexed by L-R scale using EMHMM was marginally associated with faster RT, suggesting that human readers were indeed more efficient in processing AB/C segmentation structures due to left to right sequential word processing. These results were consistent with previous research (Huang & Li, 2020). However, in contrast to Huang et al. (2021), we did not observe a main effect of frequency order. Note that Huang et al. (2021) only used sentences with AB/C as the correct OAS segmentation as the stimuli. In a separate analysis, we found that if we only considered these sentences in our stimuli, a significant frequency order effect was found, consistent with Huang et al. (2021). Our results thus suggest that when both AB/C and A/BC segmentation structures were considered, there was no main effect of frequency order.

In contrast to humans, LLMs did not show an advantage for AB/C segmentation structure in the measure of preference for the correct segmentation structure. However, since we were not able to manipulate the parameters of Davinci-003 and GPT-3.5-TI to create different versions of the model for a by-subject analysis to be compared with human data, this result was limited to by-item analysis. When we compared LLMs and humans in a by-item analysis, no interaction effect between group and segmentation structure was found. Thus, although it remains unclear whether LLMs would demonstrate a weaker segmentation structure effect than humans in a by-subject analysis, our current results suggested that they may not differ in the segmentation structure effect.

In our human data we also observed an interaction between segmentation structure and frequency order in accuracy, OAS total gaze duration, and OAS first-fixation duration. This was consistent with previous research (Ma et al., 2014). Consistent with our hypothesis, LLMs also exhibited this interaction in the measure of preference for the correct segmentation structure, suggesting that this interaction effect observed in humans could be accounted for by statistical learning of word/token occurrence regularities from the linguistic materials that the learners were exposed to without the assumption of explicit semantic word representation or sequential word processing. Note that between the two LLMs tested here, Davinci-003 showed a marginally stronger interaction effect than GPT-3.5-TI. As compared with Davinci-003, GPT-3.5-TI had fewer parameters. Thus, number of parameters may play an important role in accounting for LLMs' behavior.

When we compared humans' and LLMs' overall performance in segmenting words in OASs, humans generally outperformed the LLMs, consistent with our hypothesis. Perhaps humans could use richer internal representations of words that involve both sensory-derived and language- or cognition-derived information (e.g., Wang et al., 2020) to facilitate resolving semantic ambiguities. However, it remains possible that an LLM with a larger number of parameters can achieve a human-level ability without explicit semantic or memory representations. Future work may examine these possibilities.

When we examined the associations between humans' and LLMs' segmentation processes across stimuli, we found that LLMs' preference for the correct segmentation structure was not correlated with any of the performance or eye movement measures in the humans data, except for a marginal correlation with eye movement pattern assessed using L-R scale. This result suggested that although LLMs demonstrate human-like behavior in language processing, there may be fundamental differences in their processing mechanisms that led to differential processing across stimuli. Understanding the similarities and differences between human and LLM language processing have important implications for ways to facilitate human-AI interaction, such as providing appropriate explanations to enhance user understanding, inducing an appropriate level of user trust, and enhancing safety of AI adoption (Hsiao & Chan, 2023). Future work may use explainable AI (XAI) methods such as GradCAM (Liu, Zhang et al., 2023a; 2003b) to visualize the internal representations of LLMs and compare them with human data to better understand the processing differences between them.

In conclusion, through comparing LLMs and humans in resolving semantic ambiguities involved in word segmentation in Chinese sentence processing, we showed that although LLMs performed worse than humans in general, they showed a similar interaction between segmentation structure and word frequency order effects. This result suggested that this interaction observed in humans could be accounted for by statistical learning of word/token occurrence regularities in LLMs without assuming an explicit semantic representation of words. Nevertheless, it remains unclear whether LLMs could account for the segmentation structure effect observed in human data, which is related to humans' left-to-right sequential processing of words in a sentence. Also, LLMs' response to the stimuli was not correlated with any human performance or eye movement measure, suggesting differences in the underlying processing mechanisms. Future work will use XAI methods to better understand these processing differences between humans and LLMs to enhance human-LLM mutual understanding and facilitate LLM adoption.

Acknowledgements

We are grateful to Research Grant Council of Hong Kong (GRF #17608621). This work was also supported by a Strategic Research Grant from City University of Hong Kong (Project No. 7005995).

References

An, J., & Hsiao, J. H. (2021). Modulation of mood on eye movement and face recognition performance. *Emotion*, *21*(3), 617-630.

Bedny, M., & Caramazza, A. (2011). Perception, action, and word meanings in the human brain: the case from action verbs. *Ann. N. Y. Acad. Sci.*, *1224*(1), 81-95.

Chan, F. H. F., Barry, T. J., Chan, A. B., & Hsiao, J. H. (2020). Understanding visual attention to face emotions in social anxiety using hidden Markov models. *Cog. Emot.*, *34*(8), 1704-1710.

Chan, C. Y. H., Chan, A. B., Lee, T. M. C., & Hsiao, J. H. (2018). Eye-movement patterns in face recognition are associated with cognitive decline in older adults. *Psychon. Bull. Rev.*, *25*(6), 2200–2207.

Chan, F. H. F., Suen, H., Hsiao, J. H., Chan, A. B., & Barry, T. J. (2020). Interpretation biases and visual attention in the processing of ambiguous information in chronic pain. *Eur. J. Pain*, *24*(7), 1242-1256.

Cho, V., Hsiao, J. H., Chan, A. B., Ngo, H., King, N. M., & Anthonappa, R. (2022a). Understanding children's attention to dental caries through eye-tracking. *Caries Res.*, *56*(2), 129-137.

Cho, V., Hsiao, J. H., Chan, A. B., Ngo, H., King, N., & Anthonappa, R. (2022b). Eye movement analysis of children's attention for midline diastema. *Sci. Rep.*, *12*, 7462.

Cho, V., Hsiao, J. H., Chan, A. B., Ngo, H., King, N. M., & Anthonappa, R. (2022c). Understanding children's attention to traumatic dental injuries using eye-tracking. *Dent. Traumatol.*, *38*(5), 410-416.

Chuk, T., Chan, A. B., & Hsiao, J. H. (2014). Understanding eye movements in face recognition using hidden Markov models. *J. Vision*, *14*(11), 8.

Chuk, T., Chan, A. B., Shimojo, S., & Hsiao, J. H. (2020). Eye movement analysis with switching hidden Markov models. *Behav. Res. Methods*, *52*(3), 1026-1043.

Institute of Linguistics, Academia Sinica. (2005). *Word List with Accumulated Word Frequency in Sinica Corpus*.

Hsiao, J. H., An, J., Zheng, Y., & Chan, A. B. (2021). Do portrait artists have enhanced face processing abilities? Evidence from hidden Markov modeling of eye movements. *Cognition*, *211*, 104616.

Hsiao, J. H., & Chan, A. B. (2023). Towards the next generation explainable AI that promotes AI-human mutual understanding. *NeurIPS XAIA 2023*.

Hsiao, J. H, Lan, H., Zheng, Y., & Chan, A. B. (2021). Eye Movement analysis with Hidden Markov Models (EMHMM) with co-clustering. *Behav. Res. Methods*, *53*, 2473-2486.

Huang, L., & Li, X. (2020). Early, but not overwhelming: the effect of prior context on segmenting overlapping ambiguous string when reading Chinese. *Q. J. Exp. Psychol.*, *73*(9), 1382-1395.

Huang, L., Staub, A., & Li, X. (2021). Prior context influences lexical competition when segmenting Chinese overlapping ambiguous strings. *J. Mem. Lang.*, *118*, 104218.

Kalyan, K. S. (2024). A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Nat. Lang. Process. J.*, *6*, 100048.

Liao, W., Li, S. T. K., & Hsiao, J. H. (2022). Music reading experience modulates eye movement pattern in English reading but not in Chinese reading. *Sci. Rep.*, *12*, 9144.

Liu, G., Zhang, J., Chan, A. B., & Hsiao, J. (2023a). Human attention-guided explainable AI for object detection. *Proceedings of the 45tth Annual Meeting of the Cognitive Science Society*, Cognitive Science Society.

Liu, G., Zhang, J., Chan, A. B., & Hsiao, J. H. (2023b). Human Attention-Guided Explainable Artificial Intelligence for Computer Vision Models. *arXiv preprint arXiv:2305.03601*.

Liu, A., Wu, Z., Michael, J., Suhr, A., West, P., Koller, A., Swayamdipta, S., Smith, N. A., & Choi, Y. (2023). We're afraid language models aren't modeling amibgutiy. *arXiv preprint arXiv:2304.14399v2*.

Lo, Y. Y., Teng, X. Liao, W., Hsidao, J. (2023). Bilingual students' test-taking strategies in content subject assessments. Poster presentation at *the 45th Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

Ma, G., Li, X., & Rayner, K. (2014). Word segmentation of overlapping ambiguous strings during Chinese reading. *J. Exp. Psychol.-Hum. Percept. Perform.*, *40*(3), 1046-1059.

Perfetti, C. A. (1985). *Reading ability*. New York: Oxford University Press.

Rayner, K., Angele, B., Schotter, E. R., & Bicknell, K. (2013). On the processing of canonical word order during eye fixations in reading: Do readings process transposed word previews? *Vis. Cogn.*, *21*(3), 353-381.

Schepens, J., Marx, N., & Gagl, B. (2023). Can we utilize Large Language Models (LLMs) to generate useful linguistic corpora? A case study of the word frequency effect in young German readers.

Singh, T. (2023, September 20). *GPT-3.5 Turbo Instruct: a powerful new tool for professionals*. Medium.

Wang, X., Men, W., Gao, J., Caramazza, A., & Bi, Y. (2020). Two forms of knowledge representations in the human brain. *Neuron*, *107*, 383-393.

Zhang, J., Chan, A. B., Lau, E. Y. Y., & Hsiao, J. H. (2019). Individuals with insomnia misrecognize angry faces as fearful faces while missing the eyes: An eye-tracking study. *Sleep*, *42*(2), zsy220.

Zhang, F., Loo, B. P. Y., Lan, H., Chan, A. B., & Hsiao, J. H. (2023). Jobs-housing balance and travel patterns among different occupations as revealed by hidden Markov Mixture Models: the case of Hong Kong. *Transportation*.

Zheng, Y., Ye, X., & Hsiao, J. H. (2022). Does adding video and subtitles to an audio lesson facilitate its comprehension? *Learn Instr.*, *77*, 101542.