Transformers Can Model Human Hyperprediction in Buzzer Quiz

Anonymous ACL submission

Abstract

Humans are thought to predict the next words during sentence comprehension, but under unique circumstances, they demonstrate an ability for longer coherent word sequence prediction. In this paper, we investigate whether Transformers can model such hyperprediction observed in humans during sentence processing, specifically in the context of Japanese buzzer quizzes. We conducted eye-tracking experiments where the participants read the first half of buzzer quiz questions and predicted the second half, while we modeled their reading time using the GPT-2. The results showed that the GPT-2 can partially capture human hyperprediction. When the language model was fine-tuned with quiz questions, the perplexity value decreased. Lower perplexity corresponded to higher psychometric predictive power; however, excessive data for fine-tuning led to a decrease in perplexity and the fine-tuned model exhibited a low psychometric predictive power. Overall, our findings suggest that a moderate amount of data is required for fine-tuning in order to model human hyperprediction.

1 Introduction

It is widely recognized that the probability of a word within a specific context (i.e., surprisal) affects the difficulty of processing during incremental human language comprehension (Hale, 2001; Levy, 2008). Based on this premise, researchers have compared a variety of language models in terms of how well their surprisal correlates with human reading behavior (Wilcox et al., 2020; Kuribayashi et al., 2021; Van Schijndel and Linzen, 2021).

However recent works found that this cannot be applied to very large language models, which provides a poorer fit to human reading times. Oh and Schuler (2023) argues that larger Transformer-based models 'memorize' sequences during training, and their surprisal estimates diverge from humanlike expectations.

In those studies on cognitive modeling, self-paced reading experiments and eye-movement corpora are employed to utilize data regarding human reading times (Kennedy et al., 2013; Asahara et al., 2016; Futrell et al., 2018; Goodkind and Bicknell, 2018; Yoshida et al., 2021). These corpora typically use newspaper and novel texts as material and measure the reading time required for participants to read and comprehend the text. These works have devoted much attention to understanding everyday sentence comprehension, particularly the prediction of the next word.

In such typical sentence comprehension, psycholinguistics research has emphasized humans' use of contextual information to predict the next word while reading (Kutas and Hillyard, 1984; Altmann and Kamide, 1999; Kamide et al., 2003).

However, when comprehending a sentence, humans can sometimes make predictions about the whole sentence that go beyond the next word prediction (hereafter referred to as "hyperprediction"). This phenomenon requires comprehenders to anticipate not only the next word but also the structure of subsequent sentences. Although hyperprediction is an important aspect of human prediction in sentence processing, it has received limited attention in modeling research.

In this paper, we aim to fill this gap by evaluating the language models' capacity to model human predictive processes, particularly in tasks emphasizing hyperprediction. Specifically, we investigate hyperprediction in the context of buzzer quiz. Buzzer quiz is a popular type of quiz game (Tokuhisa, 2012), and buzzer quiz players are known to engage in this predictive process (Izawa, 2021). By investigating hyperprediction, a critical aspect of human predictive ability, we seek to provide insights into the degree to which language models resemble human predictive ability in sentence processing, not just the next word, but the entire sentence structure.

In summary, our key contributions are as fol-

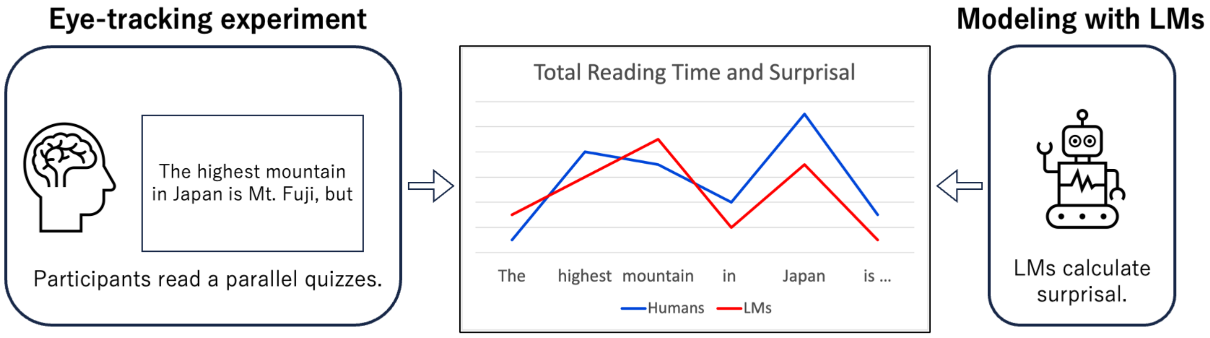

Figure 1: The process of the experiment. Human total reading time measured in the eye-tracking experiment was modeled with surprisal computed by pre-trained GPT-2 and fine-tuned GPT-2.

lows:

- This paper studies data collected from native Japanese speakers, which complements most studies using data collected in western languages.
- Our results demonstrate that the GPT-2 can partially model human hyperprediction to some extent.
- Analyses on fine-tuning reveal that fine-tuned GPT-2 can model human hyperprediction more accurately.

2 Related work

2.1 Prediction in human sentence processing

Psycholinguistics research spanning several decades has consistently suggested that humans engage in predictive processes while comprehending sentences (Ehrlich and Rayner, 1981; Kutas and Hillyard, 1984; Altmann and Kamide, 1999; Kamide et al., 2003; Pickering and Garrod, 2013; Martin, 2018). Psycholinguists have employed diverse methodologies to explore human behavior in sentence comprehension. Altmann and Kamide (1999) and Kamide et al. (2003) employed the Visual World Paradigm and revealed that humans utilize contextual cues within sentences to predict upcoming words, such as direct objects or verbs. Additionally, Kutas and Hillyard (1984) conducted EEG experiments and demonstrated that encountering a word unrelated to the context elicits a large N400 response in readers, which is associated with a semantic gap between a word and its context. Moreover, the process of next-word prediction during human sentence processing has been investigated and recent research has highlighted the necessity of the speech production system in generating lexical predictions during sentence comprehension (Martin, 2018). These studies emphasize that humans utilize the preceding context as a crucial cue for predicting upcoming words.

However, humans demonstrate the ability to predict longer sequences of words in a special situation such as in a buzzer quiz (Izawa, 2021). Skilled quiz players can answer correctly by only listening to a few words of the question sentence. In this context, they are not only required to predict the next word but also anticipate the structure of the entire sentence.

This ability to make strong predictions during sentence comprehension is a crucial aspect of sentence processing, but it has received limited attention in previous research. Therefore, this study specifically focuses on human hyperprediction.

2.2 Surprisal theory

Surprisal theory is a widely accepted concept in computational psycholinguistics, particularly in cognitive modeling research. As Eq(1) shows, surprisal is calculated as the negative logarithm of the probability of a word or sequence of words occurring in a particular context.

$$Surprisal = -\log P(word|context) \quad (1)$$

This theory proposes that the processing difficulty of a word is determined by its predictability within its preceding context (Hale, 2001; Levy, 2008; Smith and Levy, 2013). Put simply, the easier a word is to predict, the lower the cognitive load associated with it. Surprisal serves as a measure of its processing difficulty. In order to evaluate

Question	Type
サッカーのコート で、 短い方の辺 は ゴールライン ですが、 長い方の辺 は 何でしょう？ football pitch on shorter side TOPIC goal line but, longer side TOPIC what? "On a football pitch, the shorter side is the goal line, but what is the longer side?"	**easy**
南アメリカ大陸 で 最も高い山 は アコンカグア ですが、 北アメリカ大陸 で 最も高い山 は 何でしょう？ South America in the highest peak TOPIC Aconcagua but, North America in the highest peak TOPIC what? "The highest mountain in South America is Aconcagua, but what is the highest mountain in North America?"	**easy**
アメリカ合衆国 の 国の花 は バラ ですが、 メキシコ合衆国 の 国の花 は 何でしょう？ the USA 's national flower TOPIC rose but, Mexico 's national flower TOPIC what? "The national flower of the United States of America is the rose, but what is the national flower of the United Mexican States?"	**difficult**
オーストラリア の 公用語 は 英語 ですが、 オーストリア の 公用語 は 何でしょう？ Australia 's language TOPIC English but, Austria 's language TOPIC what? "The official language of Australia is English, but what is the official language of Austria?"	**difficult**

Table 1: Examples of parallel quizzes. In each question, the words in red in the first half are contrasted with those in blue in the second half. The first and second quizzes are the **easy** type of parallel quizzes, and the third quiz is the **difficult** type.

"human-like" trends of the language models, studies have been conducted to compare the surprisal calculated by language models with data obtained from humans, such as eye movement and EEG (Fossum and Levy, 2012; Smith and Levy, 2013; Frank et al., 2015; Wilcox et al., 2020; Yoshida et al., 2021).

For example, Wilcox et al. (2020); Goodkind and Bicknell (2018) compared various models by computing how well their next-word expectations predict human reading time behavior on naturalistic text corpora, and found that the less perplexity of a model, the better its psychometric predictive power.

The previous research most closely related to our work is Kuribayashi et al. (2021). They exploited the Japanese eye-track corpus BCCWJ and showed that Japanese language models with lower perplexity did not always exhibit better psychometric predictive power, which was different from English language models. This is the same trend that we reveal in this work on human hyperprediction.

Our work uses eye movement data following previous research. The surprisal calculated by the "human-like" language model is expected to correlate better with the human reading time of each word.

3 Buzzer quiz in Japanese

Buzzer quiz is a type of quiz where participants compete to answer questions quickly by buzzing in with a buzzer. In a buzzer quiz, a moderator or host reads out questions to the players. Each player is equipped with a buzzer and when players know the answer to a question, they buzz in to signal that they want to answer. The first person or team to buzz in gets the opportunity to answer the question.

While quiz players are listening to the question, they are said to predict the rest of the question sentence, not just the next word, but the entire sentence (Izawa, 2021). Typically, the players try to buzz the button even before the question is fully read.

In order to investigate human predictive processing when reading quiz questions, we experimented with *parallel quizzes*, which are typical among Japanese quizzes and where prediction is said to be important (Izawa, 2021). Parallel quizzes always have a consistent format as follows:

$$\text{For } A,\ X(A) = x_a, \text{ but what is } X(B)\ ?$$

The first half of the question sentence is the premise of the question and the second half is the main topic of the question, where B can be partially predicted from A.

Table 1 shows examples of parallel quizzes, which contrast two things in the first and second halves of the question text. In terms of the ease of predicting the second half of a question, parallel quizzes fall into two categories. The first and second questions of Table 1 are categorized as **easy** parallel quizzes, which can be answered by only listening to the first half of the question without listening to the second half. For example, the first parallel quiz on table 1 is about a football pitch. The first half of the question sentence explains the shorter edge of the pitch, then the quiz players can predict that the longer edge of the pitch will be contrasted and answer correctly (i.e., touchline) before the sentence is fully read. Skilled buzzer-quiz players can answer this kind of parallel quiz very quickly.

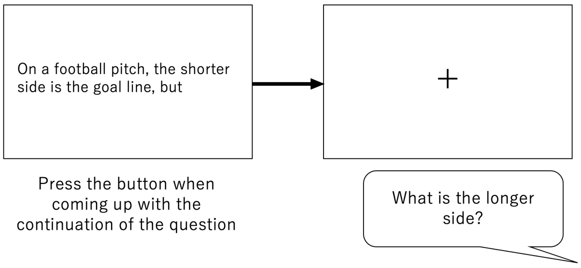

Figure 2: sentence-production task (**+predic**). Participants read the first half of a parallel quiz and predict what will follow. They orally answered the completion of the question in the second screen.

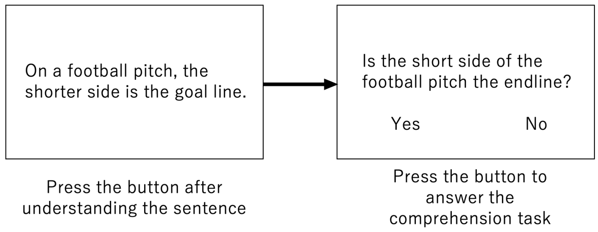

Figure 3: sentence-comprehension task (**-predic**). Participants read a sentence and answer a comprehension test on the following screen.

On the other hand, in the third **difficult** parallel quiz, the country contrasted with the word "the United States of America" is not obvious, so it is difficult to perfectly predict the second half of the question.[1]

4 Experiment

Figure 1 illustrates the experimental procedure, wherein human reading time was measured through eye-tracking experiments. Subsequently, these data were modeled using surprisal computed by language models.

4.1 Eye-tracking experiment

We conducted an eye-tracking experiment to measure the time for reading and predicting parallel questions.

Participants We recruited 32 native Japanese speakers, aged 18 to 24. Among them, seven participants were classified as **experts** due to their previous involvement in quiz clubs during high school or university, where they regularly participated in buzzer quiz activities. The remaining 25 **novice** participants had no prior experience with such activities.

Before the experiment, each participant received detailed information about the study procedures and how their data would be used. Written consent to participate in the experiment was obtained from each participant.

Stimulus sentences In this experiment, we used parallel quiz questions as stimulus sentences. All of them were extracted from a corpus of Japanese buzzer quiz questions called JAQKET. We prepared 20 **easy** parallel quizzes with a predictable second half, and 20 **difficult** quizzes with an unpredictable second half as stimulus sentences for the experiment.[2] Additionally, 40 random quiz sentences were added as fillers.

Tasks In this experiment, participants performed two types of tasks: a sentence-production task (**+predic**) and a sentence-comprehension task (**-predic**). These two tasks were shown to the participants in a randomized order.[3] In this experiment, the total reading time (TRT) of each word on the first screen was measured.

Figure 2 illustrates the process of a sentence-production task. Participants viewed the first half of a parallel quiz on the screen. They were instructed that even though there was no set time limit, they were encouraged to press the button as quickly as possible once they hit upon a continuation for the question.[4] After pressing the button, they verbally answered on the second screen.

Figure 3 depicts the procedure of the sentence-comprehension task. The first half of the parallel quiz was displayed as a declarative sentence. The participants pressed the button after reading it and answered the comprehension test on the next screen.

4.2 Language models

The surprisal for each subword was calculated using GPT-2 (Radford et al., 2019) published by rinna (Chou and Sawada, 2021) on Huggingface. Experiments were conducted using both the pre-trained

[1]One of the quiz players who participated in our experiment told that he was able to anticipate that the United Mexican States would be contrasted with the United States of America because the only two countries known as "United States" in the world are the USA and Mexico.

[2]These questions were selected from a wide range of genres to avoid bias.

[3]Each participant read 20 question sentences in **+predic** condition and the other 20 in **-predic** condition.

[4]This replicates the situation in quiz competitions, where participants must buzz in as quickly as possible.

model[5] and fine-tuned models.

The surprisal for the ith subword w_i is calculated based on the next-word probabilities $P(w_i|w_1, ..., w_{i-1})$ computed by those language models:

$$Surprisal_i = -\log P(w_i|w_1, ..., w_{i-1}) \quad (2)$$

Pre-trained GPT-2 GPT-2 calculated the surprisal for each in the sentence utilized in the eye-tracking experiment.

Fine-tuned GPT-2 We fine-tuned the GPT-2 with parallel quizzes extracted from the following resources.

- JAQKET (Suzuki et al., 2020)
  The JAQKET corpus comprises Japanese buzzer quiz questions, originally assembled for an AI competition aimed at developing systems capable of answering such quiz questions. It contains over 15,000 questions utilized in buzzer quiz competitions for college students.

- QuizWorks [6]
  This corpus comprises 18,477 questions curated by enthusiasts of buzzer quizzes. Each question is categorized by genre and format. Questions identified as “parallel quiz” were selected for fine-tuning purposes. All the quiz questions in this corpus are available for secondary use.

- Quiz-no-Mori [7]
  This website gathers numerous buzzer quiz questions utilized in competitions. Only questions that are available for secondary use were used for fine-tuning.

From these corpora, we extracted 4,100 parallel quizzes for fine-tuning. The dataset for fine-tuning was divided into 10 splits of increasing size, ranging from 10 to 4,100 data points(10, 100, 200, 300, 500, 700, 1,000, 1,500, 2,000, 4,100).[8] For each data size, we conducted fine-tuning five times using different seed values. The epoch number in training was set to ten for each fine-tuning. For conditions with 2,000 data points or fewer, the sentences used for fine-tuning were randomly selected. Importantly, none of the questions employed in the eye-tracking experiments were included in the fine-tuning data.

4.3 Evaluation metrics

Psychometric Predictive Power (PPP): The surprisal measure serves as a commonly utilized information-theoretic complexity metric. In essence, a model’s ability to predict human reading behavior is often assessed by comparing the surprisal values computed by the model with the reading times of human participants. Higher correspondence between the trends of model-generated surprisals and human reading times indicates greater psychometric predictive power. Previous studies have evaluated the psychometric predictive power of language models by comparing the surprisal values generated by each model with human reading times.

In our eye-tracking experiment, we quantified the reading time for each character and computed the total reading time for each subword by summing the total reading times of all characters within the subword.

To examine the impact of surprisal on modeling human reading behavior, we employed a linear mixed-effects regression (Baayen et al., 2008) with the `lmer` function in the `lme4` package (Bates et al., 2014) in R (R Core Team, 2023). This model aimed to predict the total reading time (TRT) of each subword using the following formula:

$$\begin{aligned}\log(\texttt{TRT}) \sim\ & \texttt{surprisal} + \texttt{length} \\ & + \texttt{is\_first} + \texttt{is\_last} + \texttt{lineN} \\ & + \texttt{segmentN} + \texttt{log\_freq} \\ & + \texttt{prev\_length} + \texttt{log\_freq\_prev} \\ & + (1|\texttt{subject\_id}) + (1|\texttt{item\_id})\end{aligned}$$

The detailed description of each variable is provided in table 3 in the Appendix.

The regression model included the surprisal factor with other baseline factors, which were previously examined in existing studies (Asahara et al., 2016; Wilcox et al., 2020; Kuribayashi et al., 2021; Yoshida et al., 2021). Factors found to be insignificant ($p > 0.05$) for modeling reading time were excluded. The frequency (`freq`) of each subword

[5]GPT-2 used in this experiment was rinna/japanese-gpt2-medium(`https://huggingface.co/rinna/japanese-gpt2-medium`). This model is published under MIT license.

[6]`https://quiz-works.com/`

[7]`https://quiz-schedule.info/quiz_no_mori/data/data.htm`

[8]The fine-tuning process with the full dataset size (4,100 data points) required approximately 15 minutes using a single NVIDIA Tesla T4 GPU.

condition	#data points	ΔlogLik ($/10^5$)
-predic	7869	1.602
+predic	8361	**1.856**
+predic, novice	6351	1.801
+predic, expert	2010	**2.140**
+predic, easy	4579	**2.390**
+predic, difficult	3782	1.912

Table 2: PPP (i.e., ΔlogLik) for each condition of the pre-trained GPT-2. These values are the mean per-word ΔlogLik of the model on held-out test data, averaged over 10-fold cross-validation. "#data points" is the number of reading time annotations used in our experiments. Surprisal values computed from the pre-trained GPT-2 were found to more accurately model the reading time for expert participants than for novice participants. Additionally, these values more effectively modeled the reading time of easy questions as compared to difficult ones.

was calculated based on the occurrences of each token within a corpus of 14 million paragraphs, extracted from Japanese Wikipedia.

To isolate the effect of surprisal on reading time modeling, we trained a baseline regression model without including surprisal information. Following the approach outlined by Wilcox et al. (2020), we computed the mean by-segment difference of log-likelihood between the model with surprisal values and the baseline model. This metric is referred to as ΔlogLik. A ΔlogLik score of zero indicates that surprisal from a language model is ineffective at all for reading time modeling. Conversely, a high ΔlogLik score suggests that the language model's surprisal values are effective for modeling reading time, indicating a high psychometric predictive power.

Considering the low amount of data, we report mean per-word ΔlogLik of the model on held-out test data, averaged over 10-fold cross-validation as suggested by Wilcox et al. (2020).

Perplexity (PPL): In order to evaluate if fine-tuning enabled the language models to better predict the next word in parallel quizzes, we calculated the perplexity of each model. PPL is the inverse geometric mean of next-word probabilities $P(w_i|w_1, ..., w_{i-1})$ in a text that consists of N words $(w_1, w_2, ..., w_N)$, and it is a typical evaluation metric for unidirectional language models:

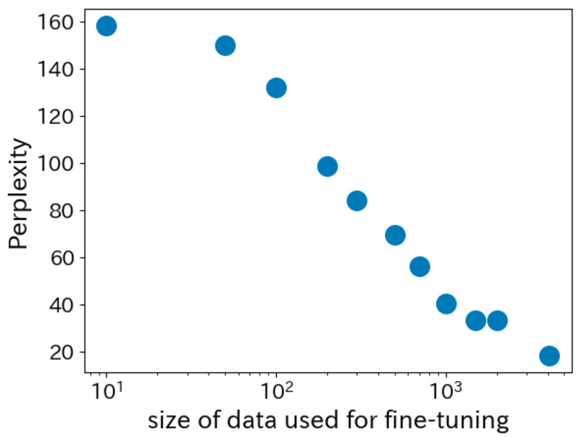

Figure 4: Relationship between the size of data used for fine-tuning (X-axis) and mean perplexity of the five fine-tuned models with different seeds (Y-axis). As the fine-tuning data set enlarges, a corresponding decrease in perplexity is observed.

$$PPL = \prod_{i=0}^{N} P(w_i|w_1, ..., w_{i-1})^{-\frac{1}{N}} \quad (3)$$

A low perplexity (PPL) suggests that the language model effectively anticipates the next word based on its contextual information. The goal of training and fine-tuning language models is to minimize the perplexity computed by the model. In our experiments, we evaluated the perplexity of a language model using texts from the eye movement data, ensuring they do not overlap with the training dataset.

5 Results

5.1 GPT-2

Table 2 shows the psychometric predictive power (i.e., ΔlogLik) for each condition of the pre-trained GPT-2. In the +predic condition, the surprisal term was found to be significantly effective in the regression model. In the sentence-production experiment (i.e., +predic condition), the participants read the first half of parallel quiz questions, and predicted what would follow. Therefore, these findings suggest that the pre-trained language model can effectively model the reading time associated with human 'hyper-prediction' when reading a parallel quiz question.

In the +predic condition, the reading time of the expert participants from the quiz club was modeled more accurately than novice participants. As for the question difficulty, the total reading time for

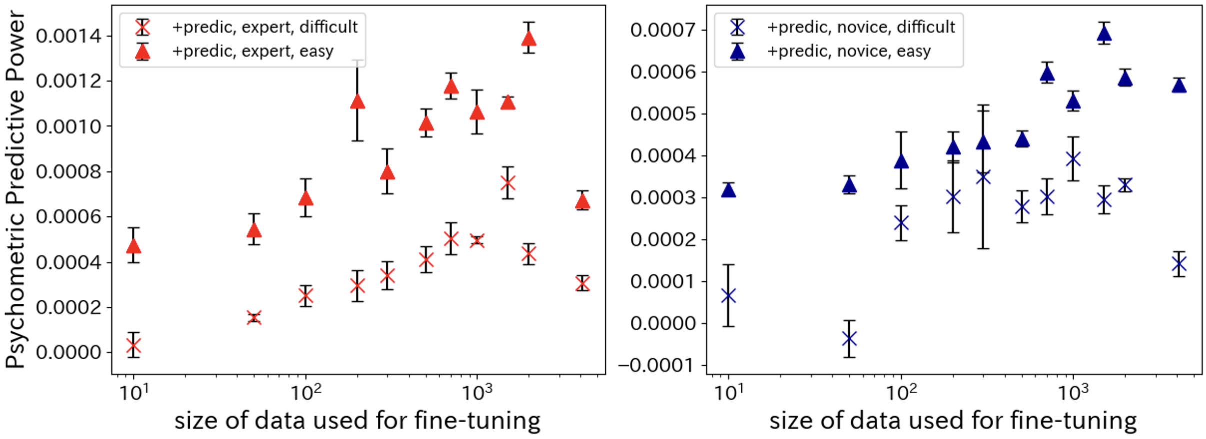

Figure 5: Relationship between the size of data used for fine-tuning (X-axis) and psychometric predictive power, i.e., ΔlogLik (Y-axis). Error bars are standard errors of by-fold mean ΔlogLik per token, using 10-fold cross-validation for five fine-tuned models with different seeds.

each subword was better modeled in easy parallel quiz questions (+predic, easy condition) than in difficult ones (+predic, difficult condition).

The results show that surprisal has more predictive power on human reading times in a condition, where one has to think of possible continuations, as compared to a baseline condition.

5.2 Fine-tuned GPT-2

Fig 5 illustrates the relationship between the size of the dataset used for fine-tuning and psychometric predictive power (ΔlogLik) of language models in +predic condition (i.e., sentence-production experiment). Each point represents a language model, with the Y-axis indicating the model's psychometric predictive power (higher scores indicate better performance) and the X-axis indicating the size of the dataset. The number of data points used for fine-tuning ranged from 10 to 4,100: 10, 100, 200, 300, 500, 700, 1,000, 1,500, 2,000, and 4,100.

Blue points represent the modeling of the reading time for novice participants, while red points represent expert participants.

As Fig 4 shows, the perplexity tended to decrease as the number of data used for fine-tuning increased.

Novice participants Language models fine-tuned with parallel quiz questions exhibited higher psychometric predictive power values than the pre-trained model. Increasing the number of data used for fine-tuning resulted in a smaller increase in psychometric predictive power.

The maximum value of psychometric predictive power was achieved with the language model fine-tuned with 1,500 sentences in the +predic, novice, easy condition and 1,000 sentences in the +predic, novice, difficult condition.

Expert participants The highest psychometric predictive power for the fine-tuned model, regardless of the number of data points used, was observed when expert participants read easy types of parallel quizzes (i.e., +predic, expert, easy condition).

In both easy and difficult conditions, the psychometric predictive power of fine-tuned models increased with the number of data points used for fine-tuning. The maximum psychometric predictive power was reached at 2,000 (+predic, expert, easy condition) or 1,500 data points (+predic, expert, difficult condition); however, beyond this threshold, a sharp decrease in psychometric predictive power was observed. Interestingly, across all four conditions, the peak psychometric predictive power did not coincide with the maximum quantity of training data.

6 Discussion

In this study, we focused on a phenomenon defined as hyperprediction, where humans are thought to predict not just the immediate next word, as is typically assumed during sentence comprehension, but also longer sequences of words and overall sentence structure. We utilized cognitive modeling techniques to examine if language models can capture this particular aspect of human prediction processing ability.

The pre-trained GPT-2 demonstrated its highest psychometric predictive power in the +predic, expert, easy condition, where human hyperprediction was expected to be most prominent. Conversely, it exhibited lower scores in the novice and difficult conditions, where hyperprediction was more challenging. Our findings suggest that even the pre-trained GPT-2 can partially capture human hyperprediction.

The surprisal from GPT-2 correlates better with the reading times of experts rather than novices, and with the +predic condition over the -predic condition. We consider that this result potentially implies the following: These results suggest that the language processing of GPT-2 aligns more with the hyperprediction capabilities of experts, who excel at predicting longer word sequences, rather than the prediction processing of average humans during normal reading. This may also suggest that the reason language models such as GPT-2 don't replicate the average human behavior is that, at least in some instances, they emulate expert behavior.

The fine-tuned models exhibited the highest psychometric predictive power in the +predic, expert, easy condition. This condition, characterized by participants' familiarity with parallel quizzes and their ease in making predictions, can be considered to reflect human hyperprediction. Language models demonstrated an ability to capture this aspect of human sentence processing.

As Fig 4 shows, the process of fine-tuning resulted in a decrease in perplexity, indicating that language models became more adept at predicting the next word in parallel quizzes. Specifically, when fine-tuned with 1,500 or 2,000 parallel quiz sentences or less, lower perplexity corresponded to higher psychometric predictive power, suggesting improved model performance.

However, The GPT-2 model fine-tuned with the most data did not necessarily exhibit the highest psychometric predictive power value. This could be attributed to the excessive data causing the model's surprisal to the sentence to decrease excessively. Consequently, the model may have failed to prioritize important words that typically require longer human reading time. This trend aligns with previous findings in Japanese language modeling research (Kuribayashi et al., 2021), which argue that lower perplexity does not always equate to human-like performance. A similar trend has been reported by Oh and Schuler (2023). They revealed that very large language models underestimated human processing difficulty. Our results align with these assertions.

7 Conclusion

This study investigated human hyperprediction in buzzer quizzes. Human hyperprediction during sentence processing involves not only predicting the next word, but also longer sequences of words and the overall structure of the sentence, which distinguishes it from regular prediction processing in sentence comprehension. In this study, we conducted experiments to test whether language models can capture this particular aspect of human predictive processing ability.

Our results showed that the pre-trained GPT-2 partially modeled human reading time while reading parallel quizzes, which suggested that language models can indeed capture aspects of human hyperprediction.

Furthermore, language models fine-tuned with parallel quizzes modeled human hyperprediction in buzzer quizzes better than the pre-trained GPT-2. Specifically, the highest predictive power was observed in conditions where hyperprediction would be most prominent (i.e., +predic, expert, and easy condition). Notably, fine-tuning resulted in a significant increase in predictive power values. However, excessive fine-tuning data (exceeding 1,500 or 2,000 data points) led to a decrease in perplexity and subsequently to reduced psychometric predictive power. This trend aligns with findings reported in previous work (Kuribayashi et al., 2021). Overall, our findings suggest that a moderate amount of data is required for fine-tuning in order to model human hyperprediction.

Limitations

Our study focused on Japanese parallel quizzes and employed an eye-tracking experiment to measure the total reading time for each subword in parallel quiz questions. However, in buzzer quiz competitions, questions are typically orally read aloud. Players utilize intonation and prominence cues to consider the answer to the quiz, particularly in parallel quizzes where the moderator emphasizes the contrasted words in the first half of the question. Skilled players exploit such phonological information to anticipate the answer and buzz in as quickly as possible. Future research could explore incorporating these oral reading dynamics into language models. Additionally, buzzer quiz players are influenced by various factors, including game rules and competitors' scores. Factors like strict penalties for wrong answers may lead players to hesitate to buzz in unless they reach a reliable prediction for the question's continuation. Conversely, players with lower scores may adopt a more aggressive approach, buzzing in even without full certainty about the answer. These varying confidence levels in predicting subsequent question text may differ from the prediction in the simplified situation of our eye-tracking experiment. Future studies can further explore these nuanced factors to gain a comprehensive understanding of quiz players' hyperprediction and the language model's ability to capture such hyperprediction.

Additionally, this eye-tracking experiment recruited a relatively small number of expert participants. There are 40 target items and 40 filler items, and given that the sentences are short, a total of 32 participants were few.

As for the statistical analysis, surprisal value was calculated for each subword. The GPT-2 tokenizer utilized in our experiment was trained using the Byte Pair Encoding (BPE) method. Consequently, since Japanese language is not written with a space between words, subwords that include a word boundary exist, resulting in reading time analyses based on subwords rather than individual words. For future work, training a tokenizer using a method that does not contain word boundaries within a single subword could allow for more cognitively valid analyses.

Ethical considerations

The eye-track experiment conducted in our work was approved by the research ethics committee of the university.

Buzzer quiz is a game of knowledge where participants may feel defeated if they are unable to answer a question. Prior to conducting the eye-tracking experiment, we emphasized to participants that the purpose of the experiment was not to assess their knowledge level. We made efforts to ensure that participants felt comfortable and performed naturally, without undue stress or pressure.

The data collected in this experiment included the timing of participants' button presses and the reading time of each word, calculated from their gaze location on the screen. These data were anonymized by assigning a random subject ID to each participant, thereby ensuring the separation of personal information from experimental data.

We aimed to ensure fair payment. As mentioned in the paper, our participants were recruited from the university and received compensation of 1,000 yen for their one-hour participation in the experiment. The compensation amount was determined following the university's guidelines.

Furthermore, in line with the ACL 2023 Policy on AI Writing Assistance, we utilized ChatGPT by OpenAI and Grammarly for writing assistance.

References

Gerry TM Altmann and Yuki Kamide. 1999. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3):247–264.

Masayuki Asahara, Hajime Ono, and Edson T. Miyamoto. 2016. Reading-time annotations for "Balanced Corpus of Contemporary Written Japanese". In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 684–694, Osaka, Japan. The COLING 2016 Organizing Committee.

R Harald Baayen, Douglas J Davidson, and Douglas M Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2014. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Tennu Chou and Kei Sawada. 2021. Publishing pretrained gpt-2 in japanese natural language processing. *The Japanese Society for Artificial Intelligence, SLUD*, 93:169–170.

Susan F Ehrlich and Keith Rayner. 1981. Contextual effects on word perception and eye movements during reading. *Journal of verbal learning and verbal behavior*, 20(6):641–655.

Victoria Fossum and Roger Levy. 2012. Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of the 3rd workshop on cognitive modeling and computational linguistics (CMCL 2012)*, pages 61–69.

Stefan L. Frank, Leun J. Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The erp response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11.

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven Piantadosi, and Evelina Fedorenko. 2018. The natural stories corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)*, pages 10–18.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Takushi Izawa. 2021. *Decomposition of Quiz Strategy*. Asahi Shimbun Publications.

Yuki Kamide, Gerry TM Altmann, and Sarah L Haywood. 2003. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and language*, 49(1):133–156.

Alan Kennedy, Joël Pynte, Wayne S Murray, and Shirley-Anne Paul. 2013. Frequency and predictability effects in the dundee corpus: An eye movement analysis. *Quarterly Journal of Experimental Psychology*, 66(3):601–618.

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower perplexity is not always human-like. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5203–5217, Online. Association for Computational Linguistics.

Marta Kutas and Steven A Hillyard. 1984. Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947):161–163.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Branzi Francesca M. Bar Moshe Martin, Clara D. 2018. Prediction is production: The missing link between language production and comprehension. *Scientific Reports*.

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Martin J Pickering and Simon Garrod. 2013. An integrated theory of language production and comprehension. *Behavioral and brain sciences*, 36(4):329–347.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Masatoshi Suzuki, Jun Suzuki, Koji Matsuda, Kyosuke Nishida, and Naoya Inoue. 2020. [jaqket: Construction of a japanese qa dataset of quizzes] jaqket: kuizu wo daizai ni shita nihon-go qa dataset no kouchiku (in japanese). *Proceedings of the Twenty-sixth Annual Meeting of the Association for Natural Language Processing*, pages 237–240.

Noriyasu Tokuhisa. 2012. *Citizen's Quiz 2.0*. Genron company limited.

Marten Van Schijndel and Tal Linzen. 2021. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive science*, 45(6):e12988.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior.

Ryo Yoshida, Hiroshi Noji, and Yohei Oseki. 2021. Modeling human sentence processing with left-corner recurrent neural network grammars. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2964–2973, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Factor name	Type	Description
surprisal	num	surprisal calculated by each language model
TRT	num	total reading time for each token
length	int	the number of characters
is_first	factor	the leftmost token within the line
is_last	factor	the rightmost token within the line
lineN	int	the serial number of the line where the token is displayed
segmentN	int	the serial number of the token within the line
log_freq	num	log of the frequency of the token
prev_length	int	length of the previous token
prev_freq	num	log_freq of the previous token
subject_id	factor	ID assigned to each participant
item_id	factor	ID assigned to each item

Table 3: Factors used in regression models.

n_layer	24
n_embd	1024
n_head	16
n_position	1024
vocab_size	32000

Table 4: Model architecture of GPT-2 we used in our work.

Optimizer	AdamW
Learning rate	5e-05
Number of epochs	10
Dropout rate	0.1
Batch size	1

Table 5: Hyperparameters for our fine-tuning.

A Factors used in regression model

Table3 shows the description of the factors used in our regression models. The frequency of a token (used in log_freq) was calculated using 14 million paragraphs extracted from Japanese Wikipadia.

B Model architecture

The model architecture of GPT-2 we used in our work is shown in Table4. The model is available on Hugging Face. [9]

C Hyperparameters

Hyperparameters for our work are shown in Table 5, which followed default settings.

[9] https://huggingface.co/rinna/japanese-gpt2-medium