Selecting Context Clozes for Lightweight Reading Compliance

Greg A. Keim
Georgia Institute of Technology
greg.keim@gatech.edu

Michael L. Littman
Brown University
mlittman@brown.cs.edu

Abstract

We explore a novel approach to reading compliance, leveraging large language models to select inline challenges that discourage skipping during reading. This lightweight 'testing' is accomplished through automatically identified *context clozes* where the reader must supply a missing word that would be hard to guess if earlier material was skipped. Clozes are selected by scoring each word by the contrast between its likelihood with and without prior sentences as context, preferring to leave gaps where this contrast is high. We report results of an initial human-participant test that indicates this method can find clozes that have this property.

1 Introduction

Ideally, college students would complete assigned readings before class, allowing professors to lean on that shared knowledge, extending and deepening understanding rather than reteaching the textbook context during the class. However, there have been a number of studies showing that when student work is not directly checked in some way, reading compliance is low (Burchfield and Sappington, 2000; Clump et al., 2004; Connor-Greene, 2000).

An obvious approach to encouraging reading compliance is for the professor or publisher to create quizzes that confirm whether students have completed the associated reading. While such questions can aid learning, thoughtfully drawing connections between various parts of the text or encouraging deeper thinking, they also require time to create, complete and score and perhaps become less useful over time as the answers begin to circulate online.

Perusall is a social learning platform designed specifically to improve reading compliance (Johnson, 2019). The tool segments students into small groups, who can then annotate and discuss the readings online. The authors report impressive results, increasing reading compliance to as much as 90% in some cases. However, this method uses group learning and written responses, which may not be possible in all situations or desirable to all students.

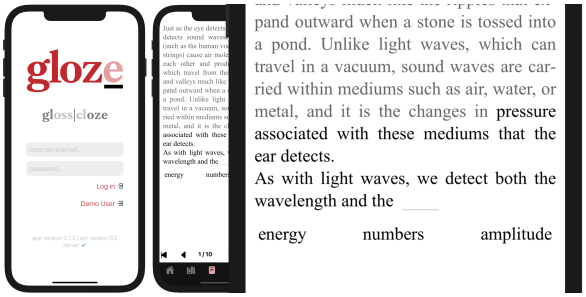

Figure 1: Gloze demonstration application, illustrating multiple choice context clozes for reading compliance.

We propose a new approach, demonstrated in a prototype application named **Gloze** (from gloss + cloze). A traditional *cloze* exercise requires a student to fill in words removed randomly or at fixed intervals from a passage. Such cloze exercises can be used to assess language proficiency, and have a long history in that literature (Alderson, 1979). In Gloze, we hope to leverage the cloze concept to increase reading compliance of long texts without time spent creating or grading external assessments. Shown in Figure 1, the method periodically requires the reader to choose the correct next word, using multiple choice with confusers to reduce the disruption of typing during reading. A key requirement of this approach is selecting challenges such that answering is easy if prior context has been read but difficult if not.

As an example of how context impacts a cloze exercise, consider the human-crafted sentence pair: *He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of* _____ . (Federmeier and Kutas, 1999). Note that the answer (*football*) is clear with the context of the first sentence (assuming familiarity with the sport), but that with only the partial second sentence, the answer is ambiguous. We define this particular cloze formulation as a *context cloze*,

Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), pages 167 - 172

where presence of prior context has an outsize (perhaps even opposite) impact relative to the immediate context. LAMBADA (Paperno et al., 2016) leverages a similar framing, though with human-computer roles reversed, to evaluate language understanding in large language models (LLMs). A test set is selected by having humans perform cloze exercises with and without broader context, selecting clozes that are easy with context and hard without.[1]

In this work, we focus on the issues of selecting context clozes with one contrasting confuser and validating that LLMs generally model the performance of human participants in this domain. Note that there are many more issues required for Gloze to be useful that are not addressed here, some of which are enumerated in the Future Work section.

All our examples and tests in this work use the text of the 685-page, freely available, anonymous college-level *Introduction to Psychology* ((Removed), 2015). However, the method could in theory be applied in any domain where confirmation that a long document has been read is important (e.g., legal agreements, safety manuals or human resources training documents).

2 Context Cloze Selection

A correlation between LLMs and human word predictions has already been demonstrated. Goldstein et al. (2020) conducted an experiment with human participants, asking them to predict each next word in a long narrative. They denoted the *predictability* of a word as the percentage of respondents that correctly generated it. Comparing human predictability scores to those from GPT-2 (Radford, 2020) on the same task, they found a strong correlation ($r = 0.79$ with a 100-word prior context). Therefore, it seems reasonable to leverage LLMs to approximate human predictability based on various contexts. In what follows, we use GPT-2 for our predictions.[2]

To choose the best context clozes with the LLM, we evaluate all words in the text, scoring each based on how the predictability changes with and without context. The reading application can use this complete weighted ordering of words to select the highest scoring cloze within some region of text. Note that this approach is not explicitly leveraging part of speech, text markup for key terms or measures of importance such as Term Frequency–Inverse Document Frequency (though our method may be implicitly finding similar "important" items, it isn't required).

After eliminating stop words, for each word in the text (a *target*), we compute this score by selecting the entire *prior* sentence and the *partial* sentence consisting of the words of the target's sentence up to the target.[3] If we define:

$t_0 = P(\textit{target}|\textit{partial})$

$t_1 = P(\textit{target}|\textit{prior+partial})$

then we prefer targets that maximize $t_1 - t_0$ (i.e., targets with high likelihood with context and low likelihood without). Note that a high-scoring target does not necessarily need to be related to the content of the chapter but simply one with the right shift in predictability.

As we aim to present these targets as cloze exercises during reading as multiple choice selections, we also consider whether there is a candidate *confuser* that actually has the opposite predictability movement. As above, we define:

$c_0 = P(\textit{confuser}|\textit{partial})$

$c_1 = P(\textit{confuser}|\textit{prior+partial})$

To select a confuser to contrast with the target from the same context, we examine the probabilities of the top 25 words in both contexts, selecting the confuser that maximizes $c_0 - c_1$ (i.e., the confuser that has the largest *decrease* in probability when context is included).

With these four next-word probabilities from GPT-2, we can define a target's score. For targets where $t_1 > t_0$ (target more likely with context), $c_0 > t_0$ (confuser more likely than target without context) and $t_1 > c_1$ (target more likely than confuser with context), we define a score $s = (c_0 - t_0) + (t_1 - c_1) + (t_1 - t_0)$. For the purposes of this work, all other words have $s = 0$.

However, we noted after initial examination of high-scoring targets that the score did not accurately capture the predictability of a student reading a textbook for a class. In particular, the student knows the subject area she is reading about, which shapes the predictability even in a partial context. To account for this observation, we added the first paragraph from Wikipedia's entry describing the field of Psychology as context in front of all prompts (Wikipedia contributors, 2021). With this

[1] State of the art systems have achieved 89.7% on this metric (Chowdhery et al., 2022).

[2] In particular, we use the 117M parameter OpenAI "gpt2" model through HuggingFace (Wolf et al., 2020).

[3] First sentences in each chapter were ignored.

Prior	Partial	Target
There are also individual differences in need for sleep.	Some people do quite well with fewer than 6	hours
If a sound occurs on your left side, the left ear will receive the sound slightly sooner than the right ear, and the sound it receives will be more intense, allowing you to quickly determine the location of the sound.	Although the distance between our two	ears
When we are awake, our brain activity is characterized by the presence of very fast beta waves.	When we first begin to fall	asleep
The BART is a computer task in which the participant pumps up a series of simulated balloons by pressing on a computer key.	With each pump the balloon appears bigger on the	screen
When you touch a hot stove and immediately pull your hand back, or when you fumble your cell phone and instinctively reach to catch it before it falls, reflexes in your spinal cord order the appropriate responses before your brain even knows what is happening.	If the central	nervous

Table 1: High-scoring example sentences from the Psychology textbook.

Prior	Partial	Target
"Checkmate," Rosaline announced with glee.	She was getting to be really good at	chess
He wanted to make his wife breakfast, but he burned piece after piece.	I couldn't believe he was ruining even the	toast
Barb loved the feel of the waves on her feet, but she hated to walk barefoot.	As a compromise, she usually wore a pair of	sandals

Table 2: CPRAG20 example sentences.

context, clozes like "The scientific ____" no longer score well, as "method" is now likely given the expanded prompt. As it also felt more disruptive to have a gap early in a sentence, a small constant was added (when $s > 0$) to favor clozes that occurred after the first few words.[4]

This procedure produces a list of all words in the textbook ranked by score (a few high-scoring samples are shown in Table 1).

3 Human-Participant Comparison

The premise of the context cloze is that we can identify places in the text where the correct next word is *unlikely* given only local context (the sentence so far) but *likely* given the prior context (the past sentence). While the probability-based scoring used to rank context clozes guarantees this condition is true for the LLM, and we know that in general LLMs approximate human language models, how can we know that this scoring method provides a reasonable model of human responses to these cloze exercises? In this section, we describe an experiment we conducted to explore this question.[5]

[4]Scores were increased by 0.5 if the partial context had at least 20 characters.

[5]This research was approved through Georgia Institute of Technology IRB, Protocol Number H21222.

3.1 Test Sets

Federmeier and Kutas (1999) measured electrical responses in the brain to understand the response to expected vs. unexpected next words in hand-built cloze exercises. They constructed 132 sentences specifically designed to have a highly likely next word response given an additional prior context sentence. More recently, this same dataset has been reused in work specifically aimed at understanding how LLMs respond to structured prompts (Ettinger, 2020), and we follow their convention in annotating the set as "CPRAG" after "Common Sense and Pragmatic Inference." In Table 2, several examples are shown to illustrate the contrast between reading both the prior context and partial sentence as opposed to just seeing the partial sentence.

While the original research involved differentiating between different types of next words (specifically, in/out of category), the measured human-participant predictability scores for these hand-built contexts can serve as a useful baseline for our own human-participant testing. Since this prior work required participants to generate the word (not choose it from a list), we use that method as well. In particular, we create two datasets with the same elements, (*prior*, *partial*, *target*), one from CPRAG and one from our context cloze method:

CPRAG20 We selected 20 examples to include in our experiment where we have average human predictability scores for both with and without context (see Table 2 for examples). The purpose of retesting this set is to validate our experimental design by replicating a prior data point.

PSYCH50 We selected 50 high-scoring context clozes from the introductory Psychology textbook (see Table 1 for examples). After programmatically scoring each word in the text, these examples

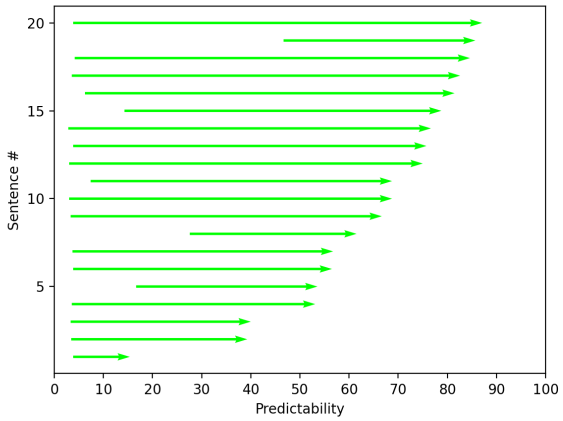

Figure 2: CPRAG20 sentences, sorted by the human predictability with context. Arrows indicate the gain in predictability from adding the prior sentence of context.

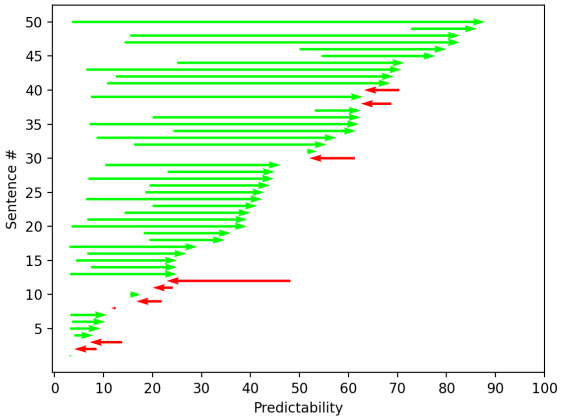

Figure 3: PSYCH50 sentences, sorted by the human predictability with context. Arrows indicate the gain or loss from adding a prior sentence context.

were selected by starting with the highest scores, skipping samples with repeated contexts, errors in sentence/word boundaries, or content that was judged might be offensive or disturbing given the absence of the full context of the chapter.[6]

3.2 Procedure

For each of the two prompt sets (CPRAG20 and PSYCH50), two sets (A and B) were created that randomly distributed with- and without-context versions of each of the prompts. Participants were randomly placed in either the A or B group of each prompt set, and then within that set they randomly responded to a subset of the prompts in a random order. In this way, no participant saw more than one form of the same prompt, everyone saw about the same number of with- and without-context prompts and any impacts from prompt order were minimized.

For each of the 70 items in CPRAG20 and PSYCH50, there were two conditions, with and without context, resulting in a total of 140 possible prompts. We tested 100 participants (online through Prolific and Qualtrics), each typing next words for 35 of these, totaling 3500 responses or about 25 answers per prompt across participants.

3.3 Results

For each target prompt attempted in the test, we calculated the predictability of the expected answer based on the percentage of respondents that typed that word. We did not correct spelling, but did remove punctuation, converted to lower case and only used the first word typed if a participant wrote the next several words.[7]

CPRAG20 Federmeier and Kutas (1999) found that human average predictability over the entire 132 sentences with context was around 74%. As selected a subset of 20 of these, we expected to get roughly the same predictability over our set. In Figure 2, predictability for both conditions of the 20 sentences are shown. Sentences are sorted by their "with context" scores, and are shown with an arrow from the without- to the with-context results. We found a 71.9% average predictability with context on our CPRAG20 subset (up from 5.7% average without), comparable to past work.

PSYCH50 As before, in Figure 3 we sort the 50 prompts by their "with context" predictability. Nine of the prompts actually decrease in predictability (shown with red leftward arrows) and some show only modest increases. However, as desired, the results do demonstrate that the method of scoring using LLMs is selecting clozes that on average show large increases in human predictability with context, from an average of 19.3% up to 47.4%, a mean absolute increase of 28.1% (with a standard deviation of 28.0).

In both the human-constructed and computer-selected cloze exercises, our testing method allowed us to confirm that some clozes have very

[6] Eighteen such examples were removed by researcher judgement.

[7] 3.7% of all responses included more than one word.

low predictability (the without-context sentences on the left of the graphs). That there is this much variation (and separation) between possible cloze prompts helps justify our focus on smart selection, so as not to frustrate the reader with clozes that are too easy or too hard.

4 Future Work

As noted earlier, the initial test presented here only addresses one piece of what an actual system would require to be successful. In future work, we want to explore larger prior context windows as well as examine fine-tuning as a replacement for the field-description paragraph. It is also critical to characterize the frequency of adequate context clozes in textbooks relative to the frequency that would be required to ensure a particular level of reading attention. In addition, many aspects of confusers need to be explored: how they impact the choice of context clozes, how various selection strategies impact compliance metrics and how to ensure LLM-generated confusers, while not the right answer by design, aren't creating confusion or demonstrating bias by being presented in a particular context. We removed some high-scoring clozes from this human-participant test, for reasons that we believe would be alleviated when the system is run in context (isolated content concerns) or through additional improvements to text processing (tokenization and filtering)–however we would need to demonstrate this is true at scale. Finally, we hope to evaluate this approach in a few classes through a mobile reading application that uses our scoring method.

5 Conclusion

We described a method for selecting context clozes to encourage reading compliance, and ran this algorithm on an introductory college textbook. We took some of the best scoring clozes and conducted a human-participant test, which confirmed our hypothesis that LLMs are a reasonable proxy for human predictability for context cloze scoring and that these clozes on average demonstrate the desired shift in predictability with and without context.

References

J. Charles Alderson. 1979. The Cloze Procedure and Proficiency in English as a Foreign Language. *TESOL Quarterly*, 13(2):219.

Colin M Burchfield and John Sappington. 2000. Compliance with required reading assignments. *Teaching of Psychology*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

M.A. Clump, H. Bauer, and C. Bradley. 2004. The Extent to which Psychology Students Read Textbooks. *Journal of Instructional Psychology*, 31(3):227–232.

Patricia A. Connor-Greene. 2000. Assessing and Promoting Student Learning: Blurring the Line between Teaching and Testing. *Teaching of Psychology*, 27(2):84–88.

Allyson Ettinger. 2020. What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Kara D. Federmeier and Marta Kutas. 1999. A Rose by Any Other Name: Long-Term Memory Structure and Sentence Processing. *Journal of Memory and Language*, 41(4):469–495.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Fanda Lora, Adeen Flinker, Sasha Devore, Werner Doyle, Daniel Friedman, Patricia Dugan, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A Norman, Orrin Devinsky, and Uri Hasson. 2020. Thinking ahead: Prediction in context as a keystone of language in humans and machines.

Steven Johnson. 2019. The Fall; and Rise, of Reading. *The Chronicle of Higher Education*, 65(31):A14–A14.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro

Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Alec Radford. 2020. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 1(May):1–7.

(Removed). 2015. *Introduction to Psychology*. University of Minnesota Libraries.

Wikipedia contributors. 2021. Psychology — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Psychology&oldid=1053798210. [Online; accessed 13-November-2021].

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Stroudsburg, PA, USA. Association for Computational Linguistics.