HumBEL: A Human-in-the-Loop Approach for Evaluating Demographic Factors of Language Models in Human-Machine Conversations

Anthony Sicilia[♭] **Jennifer C. Gates** **Malihe Alikhani**[♭]
[♭]Northeastern University
{sicilia.a, m.alikhani}@northeastern.edu gatesjenniferc@gmail.com

Abstract

While demographic factors like age and gender change the way people talk, and in particular, the way people talk to machines, there is little investigation into how large pre-trained language models (LMs) can adapt to these changes. To remedy this gap, we consider how demographic factors in LM language skills can be measured to determine compatibility with a target demographic. We suggest clinical techniques from Speech Language Pathology, which has norms for acquisition of language skills in humans. We conduct evaluation with a domain expert (i.e., a clinically licensed speech language pathologist), and also propose automated techniques to complement clinical evaluation at scale. Empirically, we focus on age, finding LM capability varies widely depending on task: GPT-3.5 mimics the ability of humans ranging from age 6-15 at tasks requiring inference, and simultaneously, outperforms a typical 21 year old at memorization. GPT-3.5 also has trouble with social language use, exhibiting less than 50% of the tested pragmatic skills. Findings affirm the importance of considering demographic alignment and conversational goals when using LMs as public-facing tools. Code, data, and a package will be available.

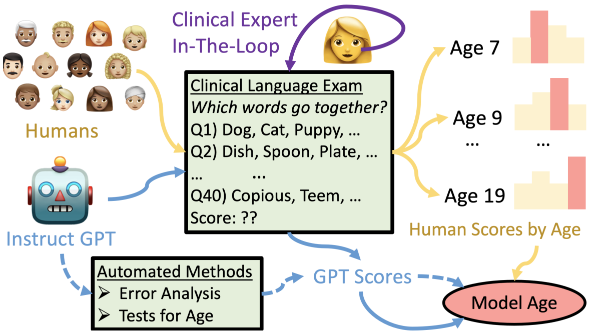

Figure 1: HumBEL uses data from human clinical exams to measure demographic factors of language models (LMs) and test alignment of LM language use with demographic groups. We propose human-in-the-loop and automated techniques.

1 Introduction

Demographic factors like age and gender impact the words we use (Sap et al., 2014; Giorgi et al., 2021) and, more broadly, the way we interact and communicate with each other (De Candia et al., 2022). Moreover, these same factors carry over influence into our conversations with machines. Age group, in particular, impacts the way we converse with household dialogue systems like Alexa (Pradhan et al., 2019), conversational agents for health information access (Harrington et al., 2022), and intelligent systems for interactive tutoring (Ogan et al., 2012). Ultimately, to effectively communicate, dialogue systems must adapt and align with the pragmatic skills, semantic understanding, and common sense of their target demographic. Despite this, there is limited work on evaluating demographic factors, and in particular, demographic alignment in human-machine conversations. To fill this gap, we propose the novel HumBEL evaluation framework,[1] which measures demographic alignment of language models (LMs) with a target user demographic for the first time. While our framework is general, we pay particular attention to modern LMs to support the rapid development of these technologies as public-facing tools.

In detail, HumBEL proposes a human-in-the-loop evaluation protocol which collaborates with a field of clinical experts (Speech Language Pathologists) that have already actively studied demographic factors in human-human communication for over 98 years (Duchan and Hewitt, 2023). These clinical experts administer language exams and compare to normative data (from large, human patient populations) to determine whether a patient aligns with a target demographic (e.g., their peers). HumBEL works by collaborating with these domain-experts to administer these same tests to a language model (LM), so key differences between LMs and human sub-populations are revealed (Figure 1). To com-

[1] Human demographic Based Evaluation of LMs

plement our human-in-the-loop clinical exams, we also propose a novel statistical test and a suite of existing statistical techniques to confirm clinician findings at scale. While HumBEL is generally applicable to any (categorical) demographic features, we focus this study on age demographics. Most importantly, our evaluation of LM alignment with different age categories can be used to examine robustness in matching conversation applications, but as a side-effect, our techniques are also able to assign a typical human age-equivalent to an LM for a specific language skill.[2]

To demonstrate HumBEL, we evaluate OpenAI's suite of GPT-3.5 models. Our key findings quantify gaps in commonsense knowledge (about noun relationships), social language use, and inference skills compared to adult human populations. Further, we find inconsistency in language skills compared to normal human development: failures in social and inferential capability are akin to error patterns of a typical 3 year old at worst or 15 year old at best, while success at recollection surpasses a typical 21 year old. Results highlight the potential for human-machine miscommunication, when the demographic factors of conversation are ignored.

Hereinafter, we introduce HumBEL, contributing:

1. (§ 2.1) protocols for evaluation of demographic factors in LMs by domain experts, using clinical exams and detailed clinician error analyses;
2. (§ 2.2) statistical tools to complement clinical techniques at scale via novel statistical tests for demographic alignment and error analysis;
3. (§ 3) comprehensive evaluation of a modern LM (GPT-3.5) using our aforementioned techniques;
4. (§ 3.3) comparison of GPT-3.5 with other modern LMs, using our statistical tools;
5. and code, data, and a python package for future researchers to easily apply our tests.[3]

2 The HumBEL Framework: Human Age Based Evaluation of Language Models

HumBEL consists of two evaluation protocols. The first (preferred) protocol describes techniques to administer a clinical exam to a LM via prompting, so that results can be carefully analyzed by a clinically licensed Speech Language Pathologist. The second describes automated alternatives, which are easier to conduct more frequently and at scale.

2.1 Clinical Evaluation by Speech Language Pathologist

In this section, we use examples from the commonly used CELF5 clinical exam (Wiig et al., 2013) to describe our protocols.[4,5,6] This test is used throughout our paper, but our ideas generalize to other common clinical tests.

2.1.1 Description of CELF5 Exam

CELF5 is composed of multiple sub-tests with 24-50 questions each. We consider the sub-tests below, which are designed to assess syntactic, semantic, and pragmatic use of language in 5-21 year olds.

1. **Word Classes (WC)** presents 3-4 words and asks test subject to identify the two words that go together best. It measures semantic knowledge and ability to apply this knowledge to determine and rank word associations.
2. **Formulated Sentences (FS)** presents 1-2 words and asks subject to provide a sentence which uses the(se) word(s). It measures syntactic and semantic correctness of the provided sentence.
3. **Recalling Sentences (RS)** presents a sentence and asks subject to repeat the sentence. It measures short-term memory and reproduction skill.
4. **Understanding Spoken Paragraphs (USP)** presents a story and asks subject questions about the story. It primarily measures recollection ability with occasional need for inference.
5. **Pragmatics Profile (PP)** analyzes social error patterns of subject, observed throughout other sub-tests as well as more targeted interactions.

2.1.2 Exam Administration via Prompting

Prompting is the standard technique in which textual output is generated from LMs. We use *prefix prompting*, in which input text is provided to the LM and the LM is sampled based on this input to complete the text. In this way, questions from the 5 discussed tests can be administered to the LM and the LM response (i.e., the text-completion) can be evaluated by the clinician with relevant observations noted for each question. Since the integrity

[2] See **Limitations**. Significant care should be taken in interpretation of LM age equivalents.

[3] Resources at: https://github.com/anthonysicilia/humbel

[4] Note, any examples of test materials provided during discussion are *adaptions* of the original materials per publishing agreement with Pearson, Inc. While different, the examples are designed to convey similar qualitative insight to the reader; e.g., the LM prompt or types of errors made by the LM.

[5] *Clinical Evaluation of Language Fundamentals, Fifth Edition, CELF-5* Copyright © 2013 NCS Pearson, Inc. Reproduced with permission. All rights reserved.

[6] *Clinical Evaluation of Language Fundamentals, Fifth Edition, CELF-5* is a trademark, in the US and/or other countries, of Pearson Education, Inc. or its affiliates(s).

`SLP`	`QA`	`Comp`
Carefully consider the following words and tell me the two words that go together best: "[W]", ...	Instruction: Carefully consider the following words and tell me... Student:	Among the words "[W]", "[X]", "[Y]", and "[Z]", the two words that go together best are

Table 1: Examples from different prompt protocols for the Word Classes test. `SLP` follows CELF5 directives exactly.[4] `QA` adds a mechanism to inform the LM of its speaker role. `Comp` re-frames as a likely seen prefix (i.e., in training). We test these and 70+ other prompt/parameter variations. See sensitivity analysis in Appendix C.

of exam results requires precise adherence to the CELF5 protocols for scoring/evaluation, we adhere to these as much as possible. We do identify two primary limitations in administering CELF5 to common LMs and provide solutions below:

1. First, some *LMs are better suited for text-completion than instruction following*, making typical administration of the test challenging. To control for performance drops induced by this, we use multiple prompt formats (see Table 1). The `SLP` protocol follows the CELF5 directives exactly, while the `QA` and `Comp` protocols should be better tailored for LMs. Sensitivity analysis (Appendix C) with 70+ additional configurations suggests prompt and parameter variations do not significantly impact LM performance.
2. Secondly, *LMs lack the ability to perceive visually and take action in an embodied setting*. Therefore, we limit the types of tests administered (i.e., those in § 2.1.1) and tailor these tests for a language-only medium when appropriate (see **Modifications**). Investigation of the impact of this choice is left for future work. Indeed, the necessity of visual/embodied stimuli to inform lexical semantics has been hypothesized (Bisk et al., 2020) and CELF5 scores may be used in the future to provide a principled answer.

2.1.3 Exam Administration via Chat

While the experimental focus is on text-completion prompts, we also conduct analysis of chat-based models, like ChatGPT. Here, we can follow CELF5 directives more precisely, but still modify tests to accommodate the limited turn-based chat medium; i.e., removing visual cues, taking scores with/without evaluation of non-verbal skills, etc.

2.2 Automation of Clinical Techniques

In this part, we describe automated techniques for two important aspects of the clinical exam: (1) qualitative analysis of errors through clinician notes and (2) determination of human demographic alignment for the LM on a task. We use the **Word Classes** test (**WC**) as an example application.

2.2.1 Data

We build a large-scale **WC** test (**WC** `large`) by combining two publicly available data sources:

1. **Word Associations**: We build associated word pairs using *cue* and *association* words from the WAX dataset (Liu et al., 2022a) collected from human annotators by presenting a *cue* and asking for spontaneous associations (with explanation). This dataset is transformed into a large-scale version of the **WC** test by randomly sampling two additional association words for each human labeled word pair and presenting the quadruple to a subject using the existing **WC** prompt protocols. All four test words (i.e., the target pair and two additional associations) are presented in random order and filtered to prevent overlap in target pairs by chance.
2. **Age Norms**: In clinical exams, human developmental standards are determined from exam score data (i.e., *age norms*) that indicate the age at which one expects the observed score in a human population. To do this automatically for new **WC** questions, we use a test-based *age-of-acquisition* (AoA) dataset (Dale and O'rourke, 1976; Brysbaert and Biemiller, 2017), which determines the AoA of 40K English words. Word AoA is determined by the age at which 50-70% of a human population *knows* the word according to a definition matching test (see Appendix A), called **Def** in experiments (§ 3). For **WC** `large`, AoA is the max AoA of the target words (i.e., the typical age at which a human can select the target pair without guessing).

Applying AoA estimates to the word association data leads to about 10K new **WC** questions with accompanying explanations and projected age norms.

2.2.2 Automated Analysis of Errors

We isolate some influential factors in typical word acquisition by humans based on discussion with a licensed Speech Language Pathologist; i.e., these question/response features were deemed useful for analyzing errors in notes during clinical exams. We limit our analysis to features that can be automatically determined.[7] The target pair features include: unordered **parts-of-speech** inferred from explanations in the WAX dataset, **relation types** from the WAX dataset, and **morphological complexity**. We also consider presence of **explanations** by GPT. Details on feature extraction are in Appendix D.

Statistical Tests In lieu of detailed notes, we propose a variety of statistical tests to determine association and impact of the various features just discussed. The χ^2-statistic provides a basic test for the association of each feature with the occurrence of an LM error. Furthermore, specific hypotheses about the impact of particular parts-of-speech, relations, and other features can be estimated using a Linear Probability Model (LPM). For example, an LPM allows us to estimate the effect size

$$\begin{aligned}&\mathbf{Pr}\{\text{LM error} \mid \text{Relation=Function}\}\\ &\quad - \mathbf{Pr}\{\text{LM error} \mid \text{Relation} \neq \text{Function}\}.\end{aligned} \tag{1}$$

while controlling for other features such as typical human age-of-acquisition for the word pair and any other features included in the model. *To summarize*, the χ^2 test lets us test the basic association between the occurrence of errors and automatically determined features, whereas the LPM lets us directly test more complicated hypotheses, e.g., "controling for AoA, does the chance of an error increase when the word pair has a functional relation?" For details on both procedures see Appendix F. Example applications are provided in later results (§ 3).

2.2.3 Automated Determination of LM Age

While we focus on age, these novel statistical tests can measure any categorical demographics.

Test Divergence We base our first test for LM age on a statistic called the *test divergence* (Sicilia and Alikhani, 2022). For an evaluation function h and language model $\texttt{LM}$ the test-divergence is:

$$\begin{aligned}\mathbf{TD}_a(\texttt{LM}) &= \mathbf{E}[|h(D) - h(\hat{D})|];\\ (D, C) &\sim \mathbb{G}_a; \quad \hat{D} \sim \texttt{LM}(C).\end{aligned} \tag{2}$$

Here, $\mathbb{G}_a$ is called the goal distribution and typically represents a distribution of human dialogues. We incorporate new dependence on the age group a, which restricts the human reference population. With this interpretation, D is a random human dialogue about the context C and $\hat{D}$ is a dialogue sampled from the language model about this same context; context can be a prompt, an image, both (for perceptually grounded models), or any other information source which grounds the dialogue. In this paper, C will correspond to a test question (or, equivalent LM prompt) in the **WC** `large` dataset and h will indicate whether the response D (or $\hat{D}$) is correct. C follows a uniform distribution over questions in **WC** `large` where AoA (§ 2.2.1) is either (1) exactly equal to a, or (2) $\leq a$. We disambiguate between these two cases throughout.

The TD Test for LM Age Granted the test-divergence as a test statistic, we are interested in the following null H_0 and alternative H_A hypotheses:

H_0 : $\texttt{LM}$ errors align with age group a
H_A : $\texttt{LM}$ errors fail to align with age group a

Thus, we grant the LM benefit of the doubt and reject the model $\texttt{LM}$ aligns with an age group if we establish evidence against this claim. Formally, we define *alignment* when a model's error patterns are within a tolerance γ: i.e., if $\mathbf{TD}_a(\texttt{LM}) \leq \gamma$. In English, this means the expected difference between the $\texttt{LM}$ performance and human (aged a) performance on each test question is no more than the tolerance γ where tolerance allows us to account for any (human) subjectivity in question responses. Then, with this, we can rewrite our hypotheses:

$$H_0 : \mathbf{TD}_a(\texttt{LM}) \leq \gamma, \; H_A : \mathbf{TD}_a(\texttt{LM}) > \gamma.$$

In turn, a test at confidence $100 \times (1 - \alpha)\%$ rejects the null if the p-value is bounded by α

$$p = \mathbf{Pr}(\widehat{T}_a - \gamma \leq T_a - \gamma \mid H_0) \leq \alpha \tag{3}$$

where $\widehat{T}_a$ is the observed estimate of $\mathbf{TD}_a(\texttt{LM})$ (i.e., an empirical average) and T_a is the r.v. representing this empirical average. For the **WC** `large` dataset, $n \cdot T_a$ is a Binomial random variable and probability under the Binomial distribution gives the p-value exactly. In other cases, the test outcome may be continuous or the test h may be learned from data similar to work by Bruni and Fernández (2017). Here, Hoeffding's or PAC type bounds can yield p-values (Shalev-Shwartz and Ben-David, 2014).

The Mean Test for LM Age As we will see in later results, the statistic/test just described will often be preferred because it incorporates information about individual question outcomes, making it more sensitive to correlation between $h(D)$ and $h(\hat{D})$. Still, we may not have access to the individual human question outcomes $h(D)$. Instead, we might only know the average outcome $\mu_a = \mathbf{E}[h(D)]$ with $D \sim \mathbb{G}_a$. Following the same logic as before, we can use this to test alignment:

[7] We use the spacy package.

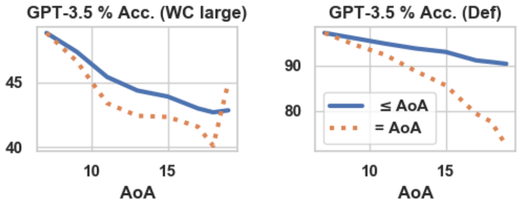

Figure 2: Accuracy of InstructGPT on **WC** `large` and **Def**.; AoA is defined in § 2.2.1. Solid line tests pairs at most the AoA. Dotted tests pairs exactly at the AoA.

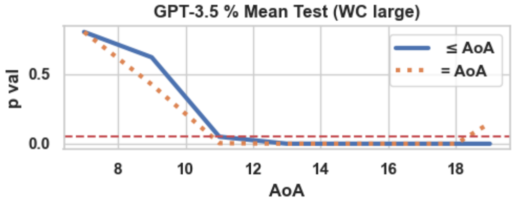

Figure 3: Vertical axis shows p-values from mean tests. Red dashed line is $\alpha = 0.05$. μ_a is estimated based on Dale and O'rourke (1976), accounting for chance and subjectivity of gold associations (see Appendix B).

$$H_0 : \mathbf{E}[R] = n \cdot \mu_a, \; H_A : \mathbf{E}[R] < n \cdot \mu_a.$$

where R is the empirical sum of correct GPT responses $\sum_i h(\hat{D}_i)$ and n is the question count. Note, this leads to a standard Binomial test.

3 Results: Applying `HumBEL` to GPT

3.1 Clinical Evaluation Results

Table 3 shows CELF5 test scores and age equivalents for InstructGPT (`text-davinci-002`) and select results for ChatGPT (`gpt-3.5-turbo-0301`).[8] We discuss qualitative clinician observations with supporting quantitative analyses, *providing italicized takeaways for conversational applications*. While this part focuses on InstructGPT, comparison to ChatGPT is provided in § 3.3. For sensitivity analysis to prompt/parameters, see Appendix C.

Modifications To adapt the **Word Classes** for language models, we remove any visual stimuli. We also include a further modified test **WC***. While official clinical evaluation stipulates the evaluator should prematurely conclude the **WC** test if 4 sequential incorrect answers are provided, this stopping rule (ceiling) is based on human development (i.e., easier words are presented earlier), which GPT may not follow. For comparison, **WC*** reports evaluation without a ceiling. Similarly, we modify the **Pragmatics Profile PP** since it measures social language capabilities which are not observable in prompt-only or turn-based chat mediums; e.g., non-verbal cues and initiative behaviors. The profile with these items removed is called **PP***.

Recollection vs. Inference *InstructGPT excels at memorization, but has trouble making inferences.* Of all tests in Table 3, Word Classes (**WC**) most requires the ability to make new inferences from existing (lexical semantic) knowledge. This is also the task that InstructGPT performs worst at, demonstrating alignment with the ability of a 6 year old. While InstructGPT was generally more successful on other tasks, the evaluating clinician observed errors in **USP** were also frequently due to trouble drawing inferences. When InstructGPT provided explanations for answers on **WC**, the clinician observed flawed or irrelevant logic in more than 59% of cases. See Table 2 for examples of inferential and other language application errors. Note, this pitfall of GPT also induces a large variation in scores (e.g., from age equivalent over 21 to under 4) which is certainly atypical of human norms. Despite some negatives, the impressive proficiency of GPT at recollection suggests *it would excel in conversational applications requiring rote information extraction*. In applications requiring inference about word meanings, *one might consider communicating the error patterns of GPT, depending on target interlocutor age and conversational goals.*

Difficult Relations *InstructGPT has more trouble with functional roles, categories, and antonyms.* On Word Classes (**WC**), the evaluating clinician identified multiple errors for each of these relation types. For functional roles, InstructGPT fails to recognize relationships like "[X] goes in [Y]" or "[X] used for [Y]". It also failed to recognize categories like "body parts", "senses" and dichotomous pairs used to describe the same concept; e.g., "brief" and "long".[4] Table 2 shows examples.

Atypical Semantic Errors *According to human developmental standards, InstructGPT understands some "hard" words better than "easy" words.* In particular, the clinician observed error patterns in semantic knowledge which were distinct from typical patterns in children. While InstructGPT failed frequently at comparatively "easy" word relations (e.g., *shirt* and *jacket*), it succeeded at "harder" relations (e.g., *copious* and *teem*).[4] In the data, this is exemplified by **WC** and the modified test **WC***. The difference in scores implies InstructGPT accumulated sequential errors early in the test on

[8] Previously accessible through the OpenAI API.

Test	`InstructGPT`	Clinician Observation
WC	Among the words "car", "water", "stroller", and "boat", the two words that go together best are "car" and "boat". Both are types of transport.	Misses functional *goes in* relation for *boat*, *water*
WC	Among the words "singing", "loving", "touching", and "tasting" the two words that go together best are "singing" and "loving." This is because both words involve using your voice and express affection.	Misses categorical *sense* relation for *touch*, *taste*
USP	Melanie greeted Miss Grace because she was happy to see her.	Missing context: Grace is old camp instructor
USP	["throw-and-chase" is] a game where one person throws a ball and the other person goes to chase it.	Not fact based. Fact-based answer is found in context

Table 2: Examples of inferential and other language application errors by InstructGPT in CELF5 exam. Explanations are provided by the evaluating clinician. Examples are adapted for publication per agreement with Pearson.[4]

model	**WC**	**WC***	**FS**	**RS**	**USP**	**PP**	**PP**$*$	**WC**	**WC***	**FS**	**RS**	**PP**
`Instruct w/ SLP`	3%	50%	94%	88%	93%	39%	48%	3:2	7:5	21:5+	21:5+	< 3
`Instruct w/ QA`	28%	50%	85%	96%	93%			5:3	7:5	12:7	21:5+	
`Instruct w/ Comp`	35%	60%	90%	100%	88%			5:11	8:10	15:1	21:5+	
`ChatGPT (0301)`	83%	83%	-	-	75%	45%	60%	14:7	14:7	-	-	< 3

Table 3: Evaluation of GPT-3.5 (`text-davinchi-002`, `turbo-0301`). (Left) Test scores reported as percent of highest possible score. (Right) CELF5 age equivalent (year:month) for scores on Left. CELF5 age equivalents are not available for **USP** or **PP***. Discussion focus in § 3 is placed on InstructGPT, while ChatGPT and other models (Table 4) are discussed in § 3.3.

model	**age range**	**accuracy**		
		@7	**@15**	**@19**
Llama-2-chat 7B	< 7	24.2	19.1	23.4
Zephyr-β 7B	< 7 to 7	36.5	27.4	21.5
Mistral 7B v0.2	7 to 11	58.7	39.6	33.8
InstructGPT	< 7 to 9	48.8	42.3	45.0
ChatGPT (0301)	15 to 19+	71.5	65.2	62.4
ChatGPT (1106)	15 to 19+	66.7	62.5	66.2
+ 3 examples	< 7 to 13*	52.4	37.5	55.4
+ 10 examples	< 7 to 19+*	51.6	44.2	68.8

Table 4: Age estimates using Mean Test on **WC** `large` with different models. μ_a is estimated as in Figure 3. Age range upper- and lowerbounds use *smaller* and *larger* estimates of μ_a, respectively, providing a *less* strict and *more* strict test of age. For both, we report the first age a at which significant difference is noted when AoA = a. Accuracy for different word AoA is also provided. Random sample of 1000 examples total are used for 7B models and ChatGPT 1106. For open-source 7B parameter models, quantization is used for inference. For ChatGPT 1106, we test impact of in-context learning using 3 or 10 random demonstrations; *Mean Test is invalid for ICL.

"easy" word relations, while still succeeding later on "hard" relations. This example hits home the necessity of considering human demographics in evaluation, since *GPT does not appear to conform to human preconceptions of how knowledge builds*. This disconnect can lead to *significant misunderstandings in conversational applications*.

Social Error Patterns *InstructGPT fails to consider context, leading to lower social capability.* In particular, the clinician observed key behaviors of InstructGPT based on the Pragmatics Profile (**PP**). InstructGPT said illogical things given the surrounding context and displayed misunderstanding of directions and goals. For example, some cases are exemplified during **WC** and **USP** in Table 2. Clinician also observed GPT provided too much information when answering questions. Note, these contextual issues are exacerbated by an LMs limited interactive capabilities; e.g., inability to use non-verbal aspects of language and initiate. We consider how these factors affect **PP** scores through **PP*** which removes these test items: the score increases considerably, but is still far from normal for humans of any age. Overall, the limited social capabilities of instruction following models "out-of-the-box" suggests *further work is needed to adapt them to (social) conversation applications.*

3.2 Automated Evaluation Results

As before, we focus in this part on InstructGPT with comparison to ChatGPT in § 3.3. Performance of InstructGPT[9] on **WC** `large` and **Def** is provided in Figure 2 with p-values from a mean test for LM age in Figure 3. We provide performance of human annotators on a 1% ($n = 108$) sample of **WC** `large` in Appendix Table 5.

Overall Performance Coarse-grained results for InstructGPT are generally consistent with the clinical evaluation results in § 3.1. Accuracy, which is equivalent to the **WC*** score in Table 3, is consistent with the clinical evaluation based on a 95%

[9] Intended answer is extracted using the first uttered test words (2 for **WC** `large` and 1 for **Def**); this was based on clinician observation on CELF5. Human evaluation of the rule on **WC** `large` ($n = 108$) also showed 100% intent recovery.

confidence interval.[10] It is notable that **WC** `large` may be more difficult, as exhibited by human disagreements (see Table 5). Overall, the general takeaways of the clinical exam can be confirmed in these coarse-grained results. For example, InstructGPT appears to succeed at the recollection task **Def**, which only requires recalling a definition, and perform worse at the inference task **WC** `large`. Also, GPT shows a spike in performance when word pair AoA is 19 (exactly), demonstrating unnatural word acquisition compared to human age standards.

Automated Determination of LM Age Based on p-values in Figure 3, we determine InstructGPT to align with ages 9- or 11-and-under for **WC** `large`, depending on whether $\mathbb{G}_a$ contains questions with word pair AoA exactly a or $\leq a$, respectively. This can be seen by excluding all ages where the means test rejects the null that GPT aligns with age group a (i.e., dipping below red line of significance). When word pair AoA is exactly 19, the means test succeeds in identifying the aforementioned "unnatural" spike in performance by correctly failing to reject the null. Overall, the means test is consistent with the clinical evaluation.

Automated Analysis of Errors In Appendix Figure 6, we visualize the influential factors on language errors discussed in § 2.2.2 and determine each has statistically significant association with the errors of InstructGPT. We also consider 6 hypotheses about these factors which were formulated through discussions with the evaluating clinician. Details are given in Appendix E. Hypotheses are tested with an LPM (see Appendix F), and results in Figure 4 confirm observations from the CELF5 exam (§ 3.1). We report each hypothesis and corresponding effect size Δ (increase in % error) below:

- **H1**: *InstructGPT has more trouble when target pairs include adverbs or adjectives ($\Delta = 3.5$).*
- **H2**: *InstructGPT has more trouble when the associated pair do not share POS ($\Delta = 3.1$).*
- **H3**: *InstructGPT has more trouble with particular relation types ($\Delta = 11$).*
- **H4**: *InstructGPT has more trouble with morphologically complex words ($\Delta = 2.3$).*
- **H5**: *GPT does worse when it explains ($\Delta = 6.2$).*
- **H6**: *InstructGPT has more trouble as word pair AoA increases ($\Delta = 0.5$; i.e., 5% from 9 to 19).*

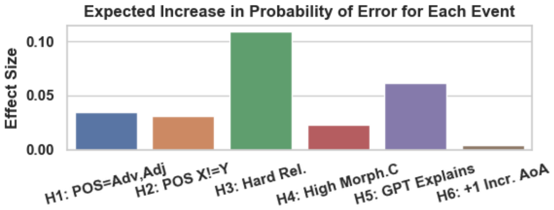

Figure 4: Expected increase in probability of InstructGPT error on **WC** `large` for different categories of word pairs. LPM estimates are significant at confidence 99% (with Bonferroni correction) except **H4**. Estimates are near true effect size for large samples (see Appendix F).

3.3 Comparison of Results with More Models

We focus on a comparison between InstructGPT and ChatGPT (`gpt-3.5-turbo-0301`) first, looking at both clinical and automated results. Then, we study age ranges and accuracy on **WC** `large` for a wider array of models, including newly released open-source models and more recent versions of ChatGPT (`gpt-3.5-turbo-1106`).

ChatGPT Clinical Results While focus is on InstructGPT, we also explored performance of a chat-based model (ChatGPT; `gpt-3.5-turbo-0301`) on CELF5. We focused on subtests **WC**, **USP**, and **PP**. These tests target aspects of inference and social language use (among other things) for which InstructGPT was poorly aligned with adult age groups. Findings (Table 3) indicate ChatGPT improves upon inference about word meanings with 23%-48% higher scores on **WC** and **WC*** compared to InstructGPT. ChatGPT also improved upon the **PP** subtest by 9%. Albeit, this score still aligns poorly with the pragmatics skills of adult humans. According to clinician notes, ChatGPT's safety features and limited chat medium (turn-based text) still severely limits its pragmatic abilities on CELF5. *It tends to avoid providing subjective opinions (even when asked), is incapable of many non-verbal aspects of social language, and does not initiate (e.g., ask questions).*

ChatGPT Automated Results We also conduct a full automated analysis on ChatGPT. The automated Mean test for LM demographic alignment shows ChatGPT aligns with ages 15-and-under when AoA = a on **WC** `large`, which again agrees with the CELF5 clinical examination. In testing, the human correctness parameter μ_a for the Mean test was increased to make the Mean test more sensitive (i.e., making a more strict/difficult test), but this was still within bounds on μ_a specified by Dale

[10] Via Hoeffding's inequality with $n = 40$ examples tested in **WC***, the two-sided interval has lower bound of 39%.

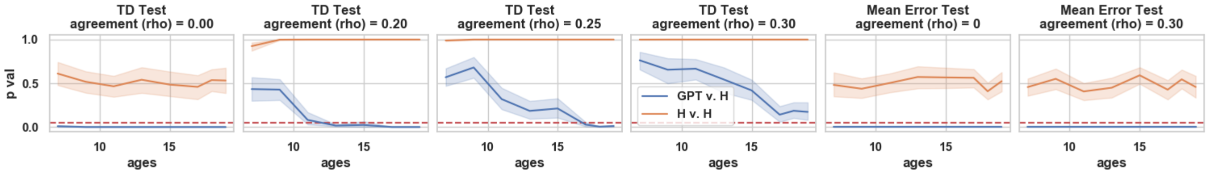

Figure 5: Bounds on p-values for TD and Means test. Red dotted line is significance level 0.05. ρ is proportion agreement.

and O'rourke (1976). The impact of changing μ_a does speak to the need for careful demographic selection, since small differences in human populations can change LM alignment. For the analysis of errors, **H1**-**H6** are consistent with results for InstructGPT, except for **H3**: ChatGPT actually does *better* when it explains, whereas InstructGPT does worse. Overall, these results echo the clinician observations that *ChatGPT has somewhat improved skill making new inferences about word meanings*.

WC large Test with More Models Automated tests provide a quick and convenient tool to quantify progress in more recent model releases. We studied three 7B parameter open-source models: Llama-2-chat (Touvron et al., 2023), Zephyr-β (Tunstall et al., 2023), and Mistral Instruct v0.2 (Jiang et al., 2023). For each, we use 4 bit quantization (Dettmers et al., 2022). We also studied newer versions of ChatGPT (gpt-3.5-turbo-1106). Results in Table 4 show open-source models tend to perform worse than InstructGPT with only Mistral Instruct v0.2 proving to be a competitive rival. Albeit, the Mistral model is still outperformed by ChatGPT. Ultimately, from these preliminary results, *we expect many takeaways for smaller (7B) open-source models to be consistent with our findings on InstructGPT*; e.g., they demonstrate poor inferences about word meanings compared to human adults. As for the newest version of ChatGPT (gpt-3.5-turbo-1106), this model offers comparable performance to its predecessor on **WC** large in terms of estimated age alignment. Some degradation on words with lower AoA is observed, but this is still consistent with human populations, and moreover, is complemented by increased performance on words with higher AoA. We leave investigation of larger open-source models to future work, but expect these to narrow the gap between closed-source and open-source technology.

WC large with In-Context Learning We also explored the impact of in-context learning (ICL) using either 3 or 10 randomly selected demonstrations. We only tested this for the newest version of ChatGPT (1106). Importantly, providing demonstrations violates aspects of the CELF5 protocol, since students are not given examples before each question. Moreover, it violates the assumptions of our statistical test, since μ_a is estimated from data on human decisions *without* demonstrations. Thus, *it is unclear to what extent the provided age range estimates for ICL are valid*, and we mark them with an asterisk $*$. Indeed, the issue of validity may also explain the inconsistency in age estimates for in-context learning. In any case, interpreting accuracy alone, it is easy to see that ICL tends to *hurt* model performance across words with varied AoA. ICL was only beneficial with 10 examples and AoA = 19. These results may call into question the benefits of ICL when making novel inferences about semantics, i.e., echoing the discussion of what ICL really "learns" in recent literature (Min et al., 2022; Chan et al., 2022; Pan et al., 2023). More thorough study of ICL, including more advanced approaches and valid statistical tests, is needed to provide confident conclusions. We leave this to future work.

3.4 Simulated Results with TD Test for Age

In the last section, we used the Means test for LM age because we did not have access to sample human question outcomes from different age groups and can only estimate the test parameter μ_a. Next, we simulate data to show the benefit of the **TD** test when access to human outcomes is available.

Setup Figure 5 shows results applying tests to LM and human samples GPT v.H as well as two (same age) human samples H v.H. Ideally, a test should fail to reject the null for all H v.H experiments and be sensitive for GPT v.H experiments, rejecting the null when appropriate. To conduct tests and study variation, we require multiple human samples. Since we only have one (used to define **WC** large), we simulate human test performance with a random variable H_i defined:

$$H_i = \begin{cases} h(\hat{D}_i) & \text{with prob. } \rho, \\ \text{Bernoulli}\left(\frac{\mu - \rho \mathbf{E}[h(\hat{D}_i)]}{1-\rho}\right) & \text{else} \end{cases} \tag{4}$$

So, we have $\mathbf{Pr}(H_i = 1) = \mu$ regardless, and ρ controls the extent to which the model `LM` and the sampled human agree. For all experiments in Figure 5, we conduct 25 trials. H_i is simulated using Eq. (4), $h(\hat{D}_i)$ is given by GPT performance on **WC** `large`, and questions for age a comprise all questions whose AoA is less than or equal to a. We estimate μ and γ from data.[11]

Failure of Means Test As the agreement parameter ρ between the sampled human and the model `LM` increases, tests using the $\mathbf{TD}$ statistic adapt appropriately, failing to reject at higher and higher ages. So, using $\mathbf{TD}$ allows us to account for context well. In comparison, the result of the means test is unchanged, demonstrating a benefit of using the $\mathbf{TD}$ statistic (when possible).

4 Related Works

Psycho-linguistic Study of LMs Other tools derived from psychology and linguistics exist across previous work on LMs. Sahu et al. (2021) use Bloom's Taxonomy (Bloom, 1956) to improve context in LM prompts for QA. Hovy and Yang (2021) develop a taxonomy of social factors to consider for LM evaluation. Cong (2022) evaluate GPT-3 using psycholinguistic tests, and Chang and Bergen (2022) use word age-of-acquisition to study development of LM word knowledge (during training) compared to humans. Comparatively, `HumBEL` is the first work to directly measure the alignment of an LM with a human sub-population, providing systematic techniques for automatic and clinician-in-the-loop evaluation of demographic factors.

LM Evaluation and Human-Likeness Evaluation strategies for generated text include metrics based on n-gram matching (Papineni et al., 2002; Lin, 2004; Vedantam et al., 2015) as well as metrics based on neural models (Sellam et al., 2020; Zhang et al., 2019; Inan et al., 2021). Bruni and Fernandez (2017); Ippolito et al. (2020); Dou et al. (2022) also propose (human or model) adversaries to discriminate between human and generated text. Our work is most related to those works considering evaluation of human-likeness (and properties thereof). For example, our techniques target commonsense knowledge, inference, and social factors as studied in a variety of works (Nair et al., 2020; Kassner and Schütze, 2020; Liu et al., 2022b). Our work builds on broad goals of evaluating human-likeness, not only in the types of tasks we test, but also in the *communication of the results to the practitioner*, presenting qualitative and quantitative results in terms of human demographic information.

NLP Tasks Many of the SLP tasks we consider have existing counterparts appearing in the NLP literature. For example, **USP** is a narrative QA task (Kočiský et al., 2018) and **WC** is, in some respects, akin to word association tests used to evaluate semantic modeling of words (Bolukbasi et al., 2016; Caliskan et al., 2017; Liu et al., 2022b). Our work extends this literature by incorporating clinician-in-the-loop feedback for the design and evaluation of these tasks, and furthermore, is the first to incorporate human demographic data for comparison of LM performance to human sub-populations.

5 Conclusion

We present `HumBEL`, which evaluates demographic factors of conversation in language models by using novel clinician-in-the-loop statistical techniques. Our framework moves beyond measuring superficial coherence of language models, instead working towards a human-explainable way to test LMs for language use and context relevance (Clark, 1996), and to compare this language use to the human sub-populations that interact with these models. For example, our techniques provide insight on the utility of LMs for inference, information-extraction, and social applications.

While the focus of this paper has been on conversational applications – e.g., understanding the communication gaps that may persist between LMs and specific human populations – a number of other applications of this framework are also realistic. For one, our tests can establish connections between human development and LMs (e.g., to build cognitive models), which may benefit diverse research communities in studying language disorders in humans. Moreover, testing alignment between LMs and human populations may be useful in evaluation of simulated worlds (Park et al., 2023) to explore how well LMs play specific roles. While more interdisciplinary work is needed, we also hope our techniques can be extended to other factors, like in cross-cultural human-machine communication.

We make the code and data of our framework publicly available, so future researchers can make use of our suite of automated statistical techniques, and protocols for clinician evaluation.

[11] μ is lower bound of a 95% Hoeffding interval around the acc. in Table 5; γ is disagreement across sim. samples of H_i.

Limitations

First and foremost, we wish to be careful about claiming our proposed techniques ascribe an intellectual age to any AI model. It is not yet clear whether the tests for human language ability we use are an appropriate "all-in-one" assessment for artificial intelligence, especially considering the vast range of specific tasks in the literature at which artificial agents can achieve super-human performance. While the tasks we study are good indicators of general language skills in humans, connections between our framework and performance generalization of AI models on untested reasoning and social language tasks are unknown. For example, factors such as overfitting, adversarial robustness, stochasticity, and prompt sensitivity can all play a new distinct role for AI models. Thus, it is better to take care and interpret our framework as designed to investigate alignment of LM language use/skills to the language use/skills of *particular* human demographic groups on *particular* language tasks. As noted, there is still significant benefit to this more careful interpretation, since our framework serves to assess model fit in conversational AI with consideration of interlocutor demographics and goals.

Second, the nature of language models produces a gap in evaluation protocols between children and these models. While we take a number of steps to alleviate these issues, there is still need for this gap to be bridged completely; i.e., so that normative age data is most accurate. Taking clinical evaluation to perceiving and embodied models is one possibility. One can also consider collecting new normative data on tasks designed for a language-only medium, or, consider using fine-grained metrics more commonly used by SLPs; e.g., preferring percentile rank among same age peers over age equivalents.

Third, we do not explicitly consider inter-annotator (i.e., inter-clinician agreement). The CELF5 exam *does already* come with estimates of inter-clinician agreement on evaluations with humans, but it is possible that working with language models produces new challenges that will ultimately invalidate this estimate. Fourth, more human data is needed to test statistics like the test divergence on real world data. Finally, our work does not explore in-depth automated analyses on other problem areas of LMs such as social language; i.e., while our clinician-in-the-loop analysis does consider pragmatics, our automated analysis focuses on inference.

Ethics Statement

The proposed approach does not explicitly evaluate societal biases inherited by language models, so any harm or bias associated with these models should be considered separately. General methods that propose to mitigate harms can help to resolve these issues, along with careful human evaluations.

For readers or users of our framework to gain access to test questions, they may need to purchase licenses from the company, university, or research lab that publishes and produces these tests. Our use of the CELF5 examination is consistent with our publishing agreement with Pearson, Inc.

Our human subject board approved our protocol. Human subjects participated voluntarily and were compensated according to the regulations approved by our human subject review board.

Acknowledgements

Thanks to Sulare Telford Rose, Ph.D., CCC-SLP for her helpful comments and feedback.

References

Heather S Battey, David R Cox, and Michelle V Jackson. 2019. On the linear in probability model for binary data. *Royal Society open science*, 6(5):190067.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience Grounds Language. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Benjamin Samuel Bloom. 1956. Taxonomy of educational objectives: The classification of educational goals. *Cognitive domain*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Elia Bruni and Raquel Fernández. 2017. Adversarial evaluation for open-domain dialogue generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 284–288, Saarbrücken, Germany. Association for Computational Linguistics.

Elia Bruni and Raquel Fernandez. 2017. Adversarial evaluation for open-domain dialogue generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 284–288.

Marc Brysbaert and Andrew Biemiller. 2017. Test-based age-of-acquisition norms for 44 thousand english word meanings. *Behavior research methods*, 49(4):1520–1523.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. 2022. Data distributional properties drive emergent in-context learning in transformers. *Advances in Neural Information Processing Systems*, 35:18878–18891.

Tyler A. Chang and Benjamin K. Bergen. 2022. Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, 10:1–16.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Yan Cong. 2022. Psycholinguistic diagnosis of language models' commonsense reasoning. *CSRR 2022*, page 17.

Edgar Dale and Joseph O'rourke. 1976. The living word vocabulary, the words we know: A national vocabulary inventory.

Sara De Candia, Gianmarco De Francisci Morales, Corrado Monti, and Francesco Bonchi. 2022. Social norms on reddit: A demographic analysis. In *Proceedings of the 14th ACM Web Science Conference 2022*, pages 139–147.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Llm.int8(): 8-bit matrix multiplication for transformers at scale.

James Diamond and William Evans. 1973. The correction for guessing. *Review of educational research*, 43(2):181–191.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A Smith, and Yejin Choi. 2022. Is gpt-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7250–7274.

Judith Felson Duchan and Lynne E Hewitt. 2023. How the charter members of asha responded to the social and political circumstances of their time. *American Journal of Speech-Language Pathology*, 32(3):1037–1049.

John E Freund, Irwin Miller, and Marylees Miller. 2004. *John E. Freund's Mathematical Statistics: With Applications*. Pearson Education India.

Salvatore Giorgi, Lyle Ungar, and H Andrew Schwartz. 2021. Characterizing social spambots by their human traits. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5148–5158.

Robin Gomila. 2021. Logistic or linear? estimating causal effects of experimental treatments on binary outcomes using regression analysis. *Journal of Experimental Psychology: General*, 150(4):700.

Christina N Harrington, Radhika Garg, Amanda Woodward, and Dimitri Williams. 2022. "it's kind of like code-switching": Black older adults' experiences with a voice assistant for health information seeking. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

William C Horrace and Ronald L Oaxaca. 2003. New wine in old bottles: A sequential estimation technique for the lpm. *Available at SSRN 383102*.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Mert Inan, Piyush Sharma, Baber Khalid, Radu Soricut, Matthew Stone, and Malihe Alikhani. 2021. Cosmic: A coherence-aware generation metric for image descriptions. *arXiv preprint arXiv:2109.05281*.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chunhua Liu, Trevor Cohn, Simon De Deyne, and Lea Frermann. 2022a. WAX: A new dataset for word association eXplanations. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 106–120, Online only. Association for Computational Linguistics.

Ruibo Liu, Ge Zhang, Xinyu Feng, and Soroush Vosoughi. 2022b. Aligning generative language models with human values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 241–252, Seattle, United States. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sathvik Nair, Mahesh Srinivasan, and Stephan Meylan. 2020. Contextualized word embeddings encode aspects of human-like word sense knowledge. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 129–141, Online. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Amy Ogan, Samantha Finkelstein, Elijah Mayfield, Claudia D'adamo, Noboru Matsuda, and Justine Cassell. 2012. " oh dear stacy!" social interaction, elaboration, and learning with teachable agents. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 39–48.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. 2023. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8298–8319, Toronto, Canada. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Alisha Pradhan, Leah Findlater, and Amanda Lazar. 2019. " phantom friend" or" just a box with information" personification and ontological categorization of smart speaker-based voice assistants by older adults. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–21.

Pritish Sahu, Michael Cogswell, Ajay Divakaran, and Sara Rutherford-Quach. 2021. Comprehension based question answering using bloom's taxonomy. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 20–28.

Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and Hansen Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1151, Doha, Qatar. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Anthony Sicilia and Malihe Alikhani. 2022. LEATHER: A framework for learning to generate human-like text in dialogue. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 30–53, Online only. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Halbert White. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, pages 817–838.

Elisabeth H Wiig, Eleanor Messing Semel, and Wayne Secord. 2013. *CELF 5: Clinical evaluation of language fundamentals*. Pearson/PsychCorp.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Determination of Word AoA

Recall, we use a test-based age-of-acquisition dataset (Dale and O'rourke, 1976; Brysbaert and Biemiller, 2017) to determine word age-of-acquisition (AoA) of 40K English words. Age is determined by U.S. K-12 grade-level and adapted to typical age equivalents (discussed later). Word grade-level is determined via multiple-choice test in which target word definitions are provided and subjects select the target amongst multiple alternatives. A word is assigned to the earliest level at which 67-80% of subjects answer correctly, equating to about 50-70% of subjects "knowing" the word at this level (accounting for chance). A word's AoA is then inferred from grade-level via typical grade-to-age mapping for U.S. K-12; i.e., age $=$ grade $+ 5$. Tests were given to U.S. (Midwest) students across a range of socio-economic and racial backgrounds with each specific word-meaning administered to about 200 subjects. As noted, besides **WC** `large`, we also test GPT-3.5 on this multiple-choice test for matching word definitions, called *Definitions* (**Def**). Alternatives are selected randomly and the prompt is: *Among the words "[W]", "[X]", "[Y]", and "[Z]", the word that most means "[Defn.]" is.*

B Estimating Human Mean Correctness

In experiments, we use a similar approach as Dale and O'rourke (1976) to estimate μ_a from word AoA, accounting for guessing and subjectivity of the task. From results of Dale and O'rourke (1976), we make a reasonable assumption that about 50-70% of humans at a particular age level *know* a word at this age level. Unless otherwise specified, we use the lower percentage, leading to a less strict test. For a human to be correct on the **WC** task, they must both know the target words *and* agree with the annotation. To compute probability for the latter, we estimate probability of agreement from Table 5 using the upperbound of a 95% Hoeffding interval for the reported % disagreement (to be less strict).[12] Then, assuming agreement and knowledge are independent, this means 38% of humans aged a will be correct based on knowledge. Finally, accounting for guessing using the score correction of Diamond and Evans (1973), this means we should expect about 47% of humans aged a to answer correctly. If the higher base correctness (70%) is assumed, μ_a is about 66%. We assume the higher base correction in Table 4, for the age lower bound, and the lower base correction otherwise. Notice, our lower and upper estimates on human mean correctness are similar to those of Dale and O'rourke (1976), but decreased to account for the subjectivity of our task.

C Prompt and Parameter Sensitivity

Although testing for the impact of various prompts and parameters is impractical when evaluation is done by a clinician, our automated version of the **WC** test provides a more practical alternative to explore the impact of these model choices. We test different parameter settings for nucleus sampling (i.e., `top_p` $\in \{0.8, 0.9, 0.95\}$) and temperature scaling (i.e., `temp` $\in \{0, 0.5, 0.7, 1\}$) as well as 11 different prompts with varying aspects of the key prompt differences highlighted in Table 1. All in all, we test differences in InstructGPT performance on a total of 77 different prompt/parameter settings on sample of 100 examples from **WC** `large`. The standard deviation in the LM scores was only 3% and a χ^2 test for independence between the settings and the error rates indicates there is no statistically significant association between the settings and the error rates. That is, performance was not significantly impacted by prompt/parameter settings.

D Feature Extraction for Error Analysis

1. **Part of Speech (POS)** While word POS is dependent on context, the explanations in the WAX dataset (Liu et al., 2022a) provide an opportunity to infer the annotator's intended POS for the word association. In particular, we can apply open-source POS parsers[13] to the annotator explanation. This strategy assumes an explanation uses a word in the same POS as intended for the word association. In case an annotator does not use the full word pair, we use "X" for unknown. Results in Figure 6 suggest GPT-3.5 error rates can vary widely based on the pairs POS, exhibiting particular association with adverbs, adjectives, and pairs having distinct POS.
2. **Relation** The WAX dataset also contains relation categories for word associations. Recall, the results of the clinical exam suggested particular relations are challenging for GPT-3.5 and the results in Figure 6 seem to suggest this as

[12] Agreement is 100 less the % disagreement. Results without the upperbound – i.e., using exact observed disagreement– are slightly different, but takeaways are generally consistent.

[13] We use the `spacy` package.

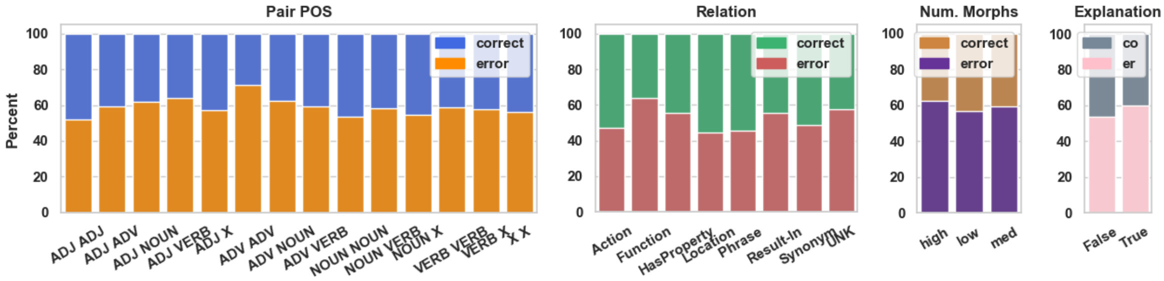

Figure 6: Proportion plot for features associated with InstructGPT errors on **WC** `large`. Association is significant at confidence 99% according to χ^2 test with Bonferroni correction. Infrequent categories not shown.

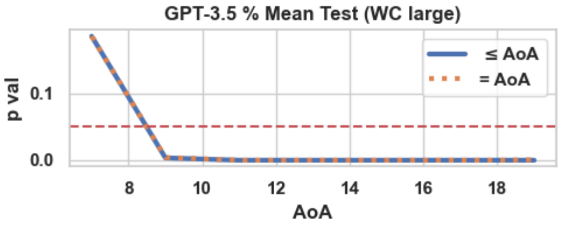

Figure 7: Results in Figure 3, re-reported without using a Hoeffding interval to estimate disagreement. Key results (i.e., lowest age estimate) differs only by a grade level.

well; e.g., as in the clinical exam, *functional* relations are hard for GPT-3.5 to identify.

3. **Morphological Complexity** We also consider Morphological Features within the Universal Dependencies framework (Nivre et al., 2016), which describe semantic and grammatical properties of words. We define *morphological complexity* as the total number of morphological features attached to at least one of the the words in the association. *High* corresponds to more than 4 features, *medium* corresponds 3-4 features, and *low* corresponds to 2 or less features. Our working assumption is that the number of features is a loose indicator of the complexity of the a word's meaning and can thus introduce challenges for GPT-3.5. The results in Figure 6 do appear to confirm this hypothesis.
4. **Explanations** Lastly, we consider if GPT-3.5 provides an (unprompted) explanation of its reasoning behind an answer. Interestingly, this occurs more times than not on the **WC** `large` dataset. While our intuition may tell us this means GPT-3.5 is more confident in the answer, the clinical evaluation actually demonstrated that GPT-3.5 often provided illogical explanations that may appear off-topic or overly complex to humans. Results in Figure 6 seem to confirm these findings, indicating that explanations typically led to worse performance at identifying associations.

E Hypothesis Selection

Below, we provide some details discussed with the evaluating clinician which led to the suite of hypotheses we test.

- **H1**: *InstructGPT has more trouble when the associated pair includes an adverb or adjective.* Clinician observations indicate trouble with modifiers in CELF5 examination. This hypothesis is confirmed in Figure 4 where we estimate a 3.5% increase in probability of error when at least one word in the pair is an adjective or adverb.
- **H2**: *InstructGPT has more trouble when the associated pair do not share POS.* Distinct POS can indicate more complex relationships across word pairs, which is a noted problem for GPT in CELF5 evaluation. This hypothesis is confirmed with a similar effect size as **H1**.
- **H3**: *InstructGPT has more trouble with particular relation types.* Building on the last hypothesis, we isolate "easy" word pair relations including {*action*, *location*, *phrase*, and *synonym* }, so the remaining "hard" word pair relations overlap with types of relations our clinician noted as difficult for GPT. Unknown relations are assumed to be hard. Results in Figure 4 confirm this hypothesis where we estimate a relatively large 11% increase in error probability for "hard" relations.
- **H4**: *InstructGPT has more trouble with morphologically complex words.* As before, assuming the complexity of a word is tied to its count of morphological features, we would expect GPT to have trouble with words having *medium* or *high* morphological feature count. We estimate an effect size similar to **H1** and **H2**.
- **H5**: *InstructGPT does worse when it explains.* Clinician evaluation on the Pragmatics checklist reveals untrustworthy, illogical explanations by GPT. Testing at scale reveals GPT has more er-

rors when it attempts to explain its reasoning with a relatively large estimated effect size of 6%.

- **H6**: *InstructGPT has more trouble as the word pair AoA increases.* While we include word pair AoA in our analysis as a potential confounder for which to control, it is also interesting to see how this variable impacts the performance of GPT. We estimate a 0.5% increase in probability of error for each unit increase in AoA; e.g., a word pair AoA of 19 would cause 5% greater chance of error than an AoA of 9.

F Overview of Statistical Tools

χ^2 Test The χ^2 test is commonly used to determine statistical association between two categorical variables (Freund et al., 2004). In our case, the two categorical variables are (1) the occurrence of a language application error by GPT and (2) one of the categorical features of the word pair discussed in § 2.2.2. The test uses a *contingency table*; i.e., a table of counts formed by letting one of the variables define the columns, the other variable define the rows, and filling each element with the number of occurrences observed for each pair of categories. Then, the test uses the statistic

$$\chi^2 = \sum\nolimits_{i=1}^{k} \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i} \tag{5}$$

where k is the number of elements in the contingency table, observed$_i$ is the observed frequency of each element of the table, and expected$_i$ is the expected frequency under the assumption that the two categorical variables are independent (i.e., the null hypothesis). Aptly, the distribution of the statistic is asymptotically χ^2 and a p-value can be calculated accordingly. We use a Bonferroni correction to control for multiple testing (i.e., across the multiple features we present as well as those not presented).

Linear Probability Model Consider a $n \times 1$ vector of dependent variables Y and a $n \times m$ matrix of independent variables X where n is the number of observations and m is a number of features for each observation. In our case, Y is a binary vector indicating the occurrence of a GPT language application error and X is a matrix ($m = 4$) with the 3 categorical features (discussed in § 2.2.2), and the last column being the word pair AoA (§ 2.2.1). With this notation, the Linear Probability Model (LPM) assumes a conditional probability model:

$$\mathbf{Pr}(Y = 1|X) = \begin{cases} 1, & X\beta > 1 \\ 0, & X\beta < 0 \\ X\beta, & \text{else} \end{cases} \tag{6}$$

Hum.	A1 ≠ A2	κ	GPT	≠ Hum.
84%	15%	0.82	56%	40%

Table 5: Sample ($n = 108$) **WC** large scores of 2 annotators aged 19+ (left) and InstructGPT (right). Annotators % disagreement and Cohen's κ is reported. GPT avg. % disagreement with annotators is reported. Annotators were prompted using the same directives as GPT; i.e., *which two words go together best?*

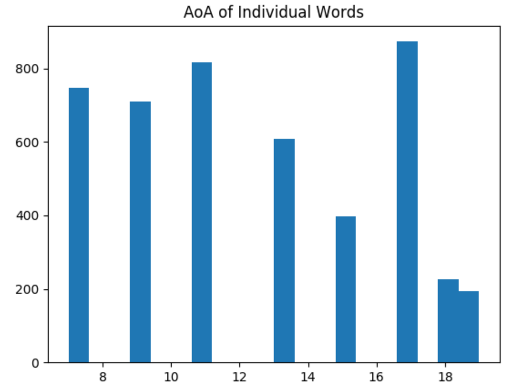

Figure 8: AoA of individual words from dataset of Dale and O'rourke (1976) used to create **WC** large.

where β is an unknown parameter vector of implied dimension. Supposing $\mathbf{Pr}(X\beta > 1) = \mathbf{Pr}(X\beta < 0) = 0$, the LPM reduces to the assumption: $\mathbf{Pr}(Y = 1|X) = X\beta$, in which case, the standard OLS estimate

$$\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y \tag{7}$$

provides a consistent estimator for the true parameter β (Horrace and Oaxaca, 2003). Techniques for heteroscedasticity (i.e., unequal variance of errors) like White's robust covariance matrix (White, 1980) can also be used to conduct hypothesis testing for significance of the coefficient estimates (Horrace and Oaxaca, 2003). We use these techniques for the coefficient estimates and statistical tests in § 3 Figure 4. As before, we employ a Bonferroni correction to control for multiple testing.

Drawbacks of LPMs Notably, the LPM has been criticized by some because it is a somewhat fragile model of the Bernoulli process governing Y (Gomila, 2021). For example, if $X\beta > 1$ or $X\beta < 0$ are probable, the interpretation of the model is unclear. Indeed, mathematically, when the presumed model is not true (e.g., when there are data such that $X\beta > 1$) the least square estimates for the LPM coefficients in Eq. (7) are biased

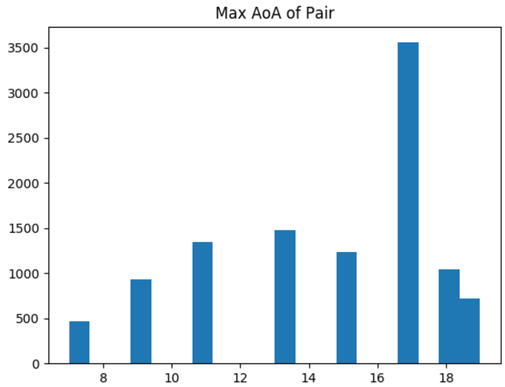

Figure 9: AoA of word pairs in **WC** `large`. Some expected accumulation in higher ages occurs (i.e., from taking a max).

(Horrace and Oaxaca, 2003). For this reason, Logistic Regression is often used instead. In our case, via standard testing procedures, one cannot refute the correctness of the LPM with data (Horrace and Oaxaca, 2003; Battey et al., 2019). Further, a logistic regression analysis led to the same takeaways as presented in the main text. Thus, we opt to show results for an LPM in the main text, since these are generally more easily interpreted (i.e., they show percent change instead of change in log odds).