

ELQA: A Corpus of Metalinguistic Questions and Answers about English

Shabnam Behzad

Georgetown University
shabnam@cs.georgetown.edu

Keisuke Sakaguchi

Tohoku University
keisuke.sakaguchi@tohoku.ac.jp

Nathan Schneider

Georgetown University
nathan.schneider@georgetown.edu

Amir Zeldes

Georgetown University
amir.zeldes@georgetown.edu

Abstract

We present ELQA, a corpus of questions and answers in and about the English language. Collected from two online forums, the >70k questions (from English learners and others) cover wide-ranging topics including grammar, meaning, fluency, and etymology. The answers include descriptions of general properties of English vocabulary and grammar as well as explanations about specific (correct and incorrect) usage examples. Unlike most NLP datasets, this corpus is *metalinguistic*—it consists of language about language. As such, it can facilitate investigations of the metalinguistic capabilities of NLU models, as well as educational applications in the language learning domain. To study this, we define a free-form question answering task on our dataset and conduct evaluations on multiple LLMs (Large Language Models) to analyze their capacity to generate metalinguistic answers.

1 Introduction

Language is so powerful that it can be reflected back on itself. Statements like “In informal usage, a *steep learning curve* means something that is difficult (and takes much effort) to learn” or “In some cases, an adjective has both *-ic* and *-ical* forms, with no difference in meaning” expressly concern linguistic inventories, structures, and behaviors. In other words, they are *metalinguistic*—they use language to discuss language (cf. Wilson, 2013). They may concern a particular instance of language use, or properties of a language or speaker in general; either way, they are metalinguistic in making linguistic phenomena the subject matter of a linguistic utterance. For the rest of this paper, the term *meta-language* is used for natural language text in which natural language is also the subject matter.

While NLP models have become powerful at *predicting* text in many settings, it remains to be seen whether such capability extends to *metalinguage*—where linguistic strings are not being deployed to

contribute to the discourse with their normal denotations, but rather, are treated as entities with linguistic properties (e.g., grammar, meaning). One way this can be explored is in a question answering framework, which requires suitable datasets, ideally based on questions that are realistic and paired with high-quality answers.

In this paper, we present a corpus of metalinguistic questions and answers about English. The corpus is collected and carefully processed from two Stack Exchange forum sites: *English Language & Usage* (ENG) and *English Language Learners* (ELL). It covers more than 70k questions on numerous topics about English such as grammar, meaning, fluency, and etymology along with answers. Our corpus, ELQA (English Language Questions and Answers), can serve as a tool to facilitate metalinguistic studies. Moreover, since questions in ELQA cover a variety of topics in English, it can be used in the educational and English language learning domains.

As the first case study of ELQA, we investigate the performance of current state-of-the-art NLP technology on free-form question answering in the English language domain. Additionally, we explore the possibility of building NLP models that can directly answer questions from language learners. We process a subset of ELQA and make it appropriate for this task. Then, we report on the results of both automatic and human evaluations using different experimental settings of T5¹ and GPT-3² models. Although most of these models achieve high ratings for well-formedness, the validity of their answers is significantly lower than that of human-authored answers, indicating that this type of metalinguistic QA task is challenging even for large language models.

Our main contributions are: 1) we release the

¹<https://github.com/google-research/t5x>

²<https://openai.com/blog/gpt-3-apps>

Dates and times: "on", "in", "at"?

Asked 9 years ago Active 3 years, 7 months ago Viewed 1k times

I'm often confused when I speak about times and dates. What is the rule for using *on*, *in*, and *at* in the following sentences?

28

- I will do it ___ Tuesday.
- We married ___ March.
- He returned ___ the same day.
- Every day ___ the same time, I walk the dog.

prepositions difference in-on-at

Share Improve this question Follow

edited Dec 18 2016 at 13:28
ColleenV
11.6k 11 43 80

asked Jan 23 2013 at 21:04
bytebuster
7,805 9 40 82

Times usually get at.

25

Everyday at the same time I take a walk.
At 3 PM, I will be having a late lunch.

✓

Days usually get on.

I will do it *on* Tuesday.
He returned *on* the same day.

↻

Months usually get *in*.

We married *in* March.

Share Improve this answer Follow

answered Jan 23 21
waiwai933
3,541 2

(a) Question

(b) Answer

Figure 1: Page screenshots from ELL Stack Exchange.

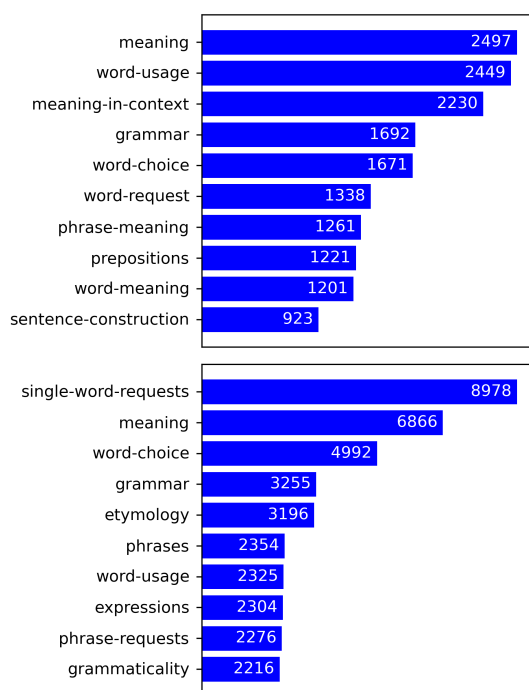


Figure 2: Frequencies of top 10 user-assigned tags in ELQA-large subcorpora: ELL (top) and ENG (bottom).

first publicly available metalinguistic QA dataset,³ focused on the English language; 2) we present a taxonomy of questions in the corpus along with analysis; and 3) we investigate to what extent LLMs are able to articulate appropriate generalizations about language in response to these questions.

2 Related Work

Stack Exchange is a network of numerous CQA sites (originally and most famously, *Stack Over-*

³<https://github.com/shabnam-b/ELQA>

flow) built on a common platform. Stack Exchange forums have been featured in a number of previous datasets (Yao et al., 2013; Hoogeveen et al., 2015; Ahmad et al., 2018; Penha et al., 2019; Campos et al., 2020; Kumar and Black, 2020; Rogers et al., 2023), including the *English* site (our ENG) along with others such as *Ask Ubuntu*, *Android*, *Gaming* and *WordPress* (dos Santos et al., 2015; Nakov et al., 2017). We focus on ENG and ELL as they concern the English language itself; we show that these datasets cover a wide range of metalinguistic questions.

Our use of these forums contrasts with previous work on metalanguage in corpora, which annotated and quantified mentions (Anderson et al., 2004; Wilson, 2010, 2011, 2012, 2017), but did not consider entire questions and answers about language. Taylor (2015) studied metalanguage in online forums, but with a focus on the usage of metalinguistic expressions of mock politeness. More recently, Bogetic (2021) published the first corpus of contemporary Slovene, Croatian and Serbian media metalanguage texts.

So far, metalanguage has not been a focus in the QA domain—ours is the first publicly available English metalinguistic QA dataset. Most QA tasks are set up to have a question and a reference document, where the objective is to find the answer based on the document (Fan et al., 2019; Kwiatkowski et al., 2019). In this paper, we explored a type of “closed-book” question answering task (Roberts et al., 2020; Khashabi et al., 2021). To the best of our knowledge, this task has not been explored to date within the realm of English language questions

ELQA-large	ELL	ENG
Total # of Qs	23,520	47,532
Total # of As	49,345	152,315
Avg. Q length	92.41	102.41
Avg. A length	158.25	137.90
Max. A score	392	581
Min. A score	-13	-28
Avg. A score	4.85	5.15
Total # of available tags	513	951
ELQA-small	ELL	ENG
Total # of Qs	6,477	14,234
Total # of As	18,389	62,744
Avg. Q length	84.21	89.25
Avg. A length	156.29	118.66
Max. A score	392	581
Min. A score	-13	-13
Avg. A score	6.63	6.73
Total # of available tags	437	823

Table 1: ELQA statistics on Qs (questions) and As (answers). To calculate average length in tokens, sequences were tokenized using SpaCy (<https://spacy.io/>).

that require significant generalization and adaptation rather than looking up facts.

3 Constructing the Dataset

We collect our data from two sites on Stack Exchange: *English Language & Usage* (ENG)⁴ and *English Language Learners* (ELL).⁵ Sample screenshots of the site are shown in Figure 1. The Stack Exchange data is publicly released under the CC-BY-SA 3.0 license. We preprocessed the data until 2021-12-06 collected from the Internet Archive⁶ to be suitable for NLP studies and release it as ELQA. Additionally, some cleanup (e.g., removing posts marked as “spam” or “offensive”) was done. Fields for each entry (question) include the title, body, user bio (if available), score (which is calculated based on up-votes and down-votes by other users), tags (user-assigned, related to the area/topic of the question), favorite count, and a list of answers. Textual content (body and user bio) is provided in two formats: HTML and plain text without HTML tags.

We release two versions of ELQA based on different preprocessing steps. In ELQA-large, we keep questions as long as they don’t include any images (HTML tag) and have an answer with a score of at least 2 (meaning at least two people other than the user posting the answer found it helpful). For ELQA-small, we applied further filtering to ensure that the data has the least amount of noise: a) questions should have a score of at least

2 (ensuring questions are clear and coherent), b) question has an answer with a score higher than 3 and c) there are no hyperlinks in at least one of the high-rated answers. The last step reduces noise and facilitates a fair comparison for the closed-book question-answering task (§4) with model-generated answers, as models cannot be expected to have access to the web to suggest valid URLs compared to humans who would search the web for appropriate resources to include in their answers.

For quality assurance, we also did a human annotation on ELQA-small. Two of the authors annotated 250 question and answer pairs for the following: 1) Is the question answerable? and 2) Does the answer fully address the question? We found 99.2% of the questions answerable and 91.8% of the answers acceptable.

Table 1 contains overall statistics on both versions. Figure 2 shows the distribution of the 10 most common tags in each of the sites. Since users assign these tags to their questions (0 to multiple), similar or near-duplicate tags are common within the collection. Some form more general and more fine-grained variants, e.g. ‘meaning’ and ‘meaning-in-context’. In addition to available user-assigned tags, we manually inspected a large subset of the data to identify salient types of questions. These are defined below and illustrated in Table 2. We then labeled 100 random questions to get a rough estimate of their frequencies (two annotators annotated these 100 samples and they agreed on 92% of cases in an overlapping subset).

- **Fluency** ($\approx 38\%$ of questions): Usually asking about a particular sentence, comparison of multiple sentences, and/or probing how an expression should be used in general. The user wants to know if X is correct, or to decide between multiple choices, which one is correct. “Correct” could mean grammatical, most natural/idiomatic, stylistically appropriate, conveying the intended meaning, etc. In Qs where options are provided by the user, there are cases in which 1) none of the choices are correct, 2) multiple choices are correct, and 3) only one is correct.
- **Form to Meaning (Interpretation)** ($\approx 19\%$ of questions): Questions such as “What does X mean?” (of an expression in general, or an encountered passage) or “What’s the difference in meaning between X and Y?”.
- **Meaning to Form (Encoding)** ($\approx 20\%$ of questions): In these questions, the user gives some

⁴<https://english.stackexchange.com/>

⁵<https://ell.stackexchange.com/>

⁶<https://archive.org/>

Question Type	Title	Body
Fluency	“On my own way vs. “in my own way”?”	Which one is correct in or own way? <blockquote> I usually help my closest friends on/in my own way. </blockquote>
Form to Meaning	Wondering what “get by” means in this context	<blockquote> He tries to get by with the least amount of work possible. </blockquote> Could you tell me what this sentence means?
Meaning to Form	Grammatically correct synonym for “level of catastrophichness”	I’m trying to say something like this: <blockquote> We have developed a strategy to numerically rate the relative level of catastrophichness of a potential hardware failure. </blockquote> Looking at a thesaurus hasn’t really helped me with this one. Can someone help me to convey this without using this ugly, incorrect grammar?
Grammatical Analysis	Should I modify a gerund using an adjective or an adverb?	I know that a gerund is a noun, so it should be modified by an adjective. However, it is also a verb form. Can I modify it by using an adverb?
Other	What is the etymology of ‘physician’	I find myself confusing ‘physician’ and ‘physicist’ occasionally. While I know what they both mean, I am a little confused as to the use of ‘physics’ in ‘physician’. How did the term ‘physician’ come to be used the way it is meant today? Lucky coincidence?

Table 2: Example posts from ELL and ENG sites for different question types. (Original post URLs and author profile URLs are all available in the Appendix.)

explanation/definition and asks for the term or for form to express it.

- **Grammatical Analysis** (*≈11% of questions*): Questions about parts of speech and other aspects of syntactic analysis. (e.g. “Is this a verb or an adjective?”; “Can an article ever go after the noun it modifies?”). Note that Fluency questions may mention grammatical terminology, but the grammatical categories are not the focus.
- **Other** (*≈10% of questions*): Any other type of question not listed above. This includes questions about pronunciation, etymology, etc.

As can be seen from the examples in Table 2, it is common for questions and answers to contain example usages, often visually distinguished with Markdown formatting (such as blockquotes, bullets, and italics) which we retain in the processed corpus markup. Examples can be incorporated into a post in a variety of ways—e.g., asking for an interpretation of one usage, as in the Form to Meaning example in Table 2, or contrasting multiple usages such as in the following question:

Did VS Have done

What is difference between the following statements: Did you tell your parents yet? Have you told your parents yet? Haven’t you told your parents yet? Are these questions correct? why do we use one over another in some cases? What is the difference in meaning?

Usage examples provided in a question may be instances that the author encountered “in the wild” (such as in a novel or film), or in a grammar book or dictionary, or they may have been constructed

by the user. Answers sometimes include examples found through a corpus search.

4 English Language Question Answering

Large language models can produce output that is fluent and (at times) informationally adequate when presented with factual questions about entities in the world (Roberts et al., 2020). But how do such models perform when asked questions about the language itself? In this section, we investigate the free-form English language question answering task.

This task has the potential to benefit educational applications for language learners. Research on NLP for educational purposes has investigated tasks such as automated grammatical error correction (Dale et al., 2012; Ng et al., 2014; Bryant et al., 2019; Wang et al., 2021, *inter alia*), question and quiz generation for language learning (Sakaguchi et al., 2013; Chinkina and Meurers, 2017; Marrese-Taylor et al., 2018; Vachev et al., 2021), and automated essay scoring (Burstein, 2003; Farag et al., 2018, *inter alia*). Nevertheless, an application that has not been taken up by the educational NLP community is free-form question answering about language. Second language learners possess a degree of metalinguistic awareness about the language they are learning, and often turn to teachers or more advanced speakers with explicit questions about vocabulary, grammar, and usage. Community Question Answering (CQA) websites such as Stack Exchange have sites for language learners’ questions and answers. These sites require consid-

	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore
GPT-3 FS	31.2	8.5	20.3	10.8	85.7
GPT-3 FT-1000	27.1	7.0	18.7	11.8	85.2
GPT-3 FT-100	25.6	6.0	17.9	9.9	85.2
T5-xxl	28.1	8.0	19.8	4.7	80.3
T5-1	21.2	6.6	17.7	4.1	78.3

Table 3: Automatic evaluation scores (percentage) for different setups. The highest value in each column is bolded.

	ROUGE-1		ROUGE-2		ROUGE-L		BLEU		BERTScore	
	ENG	ELL	ENG	ELL	ENG	ELL	ENG	ELL	ENG	ELL
GPT-3 FS	30.4	32.8	8.0	9.7	20.0	21.1	11.9	8.7	85.7	85.8
GPT-3 FT-1000	26.0	29.6	6.3	8.6	18.2	19.7	11.7	11.8	85.2	85.4
GPT-3 FT-100	24.8	28.0	5.4	7.3	17.6	18.8	9.8	10.0	85.1	85.2
T5-xxl	26.8	31.0	7.1	10.1	19.1	21.4	4.4	5.0	80.2	80.4
T5-1	20.3	23.2	5.8	8.3	17.1	19.1	3.9	4.1	78.0	79.0

Table 4: Automatic evaluation scores (percentage) for different setups broken down by site

erable effort by volunteers, and learners may have to wait for an answer—if an answer is provided at all. In fact, looking at the data from 2021-12-06 for ENG and ELL, 9% of questions have no answers.

4.1 Data

We randomly divided ELQA-small into train/test/dev splits. This resulted in 21,175 Q&A pairs in the train split and 3,107 Q&A pairs in each of the dev and test splits. Answers in these splits have a score of at least 4. If there are multiple high-rated answers to a question, we include all of them for training. Some of these questions can be answered by looking at a dictionary or vocabulary list for descriptions. But many of them are explanations in relation to particular instances of language use and require significant reasoning rather than looking up facts. Thus in this setup, we do not have any external context/reference available at evaluation time, i.e. this is a closed-book QA task.

The input for the task is *Title: [Q title] <sep> Body: [Q body]*. We use the HTML version of ELQA for this task since metalinguistic mentions are usually distinguished via formatting (e.g., blockquotes, bullets) and the ultimate goal is a system that humans can easily use to get answers to their language-related questions.

4.2 Setup

We use T5 (Raffel et al., 2020; Roberts et al., 2022) and GPT-3 (Brown et al., 2020) as our models since they have been shown to be strong baselines in other QA domains. We believe the questions in ELQA offer new challenges for the QA task since they require different types of knowledge/understanding to be able to generate answers. Addition-

ally, these questions contain noise (grammatical errors) and cases of textual metalanguage which is likely harder to comprehend for a model.

We fine-tune *T5-1* and *T5-xxl* for this task.⁷ We saved multiple checkpoints during fine-tuning and evaluated them with the interpolation of BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2020) and ROUGE (Lin, 2004) on the dev set to choose the best-performing one (checkpoint at 75k updates, hyperparameters available in Table 8 in the Appendix).

With GPT-3 we used *text-davinci-003* and experimented with both fine-tuning (FT) on 100 and 1000 samples and a few-shot (FS) setting in which the model is given a few demonstrations of the questions and answers at inference time as conditioning, but no weights are updated (Radford et al., 2019). In the FS setting, we show the model four Q&A pairs since we wanted the model to see different question types but there were also limits on the input length. To select these 4 pairs, we randomly created 5 different sets of Q&A pairs, evaluated on a subset of dev, and chose the best-performing set for the experiments (dev results available in Appendix, Table 9).

4.3 Results

4.3.1 Automatic Evaluation

Results are shown in Table 3. *GPT-3 FS* outperforms all other methods in all metrics with a large margin except for BLEU Score. We also observed that using GPT-3 in a few-shot setup worked much better than the fine-tuned version. Looking at some of the model-generated answers, we noticed that the fine-tuned model tends to generate longer an-

⁷This took 5 days with v3-8 TPU (provided by Google)

Source	C1				C2			
	Avg. on ENG	Avg. on ELL	Avg.	z	Avg. on ENG	Avg. on ELL	Total Avg.	z
Top-rated human	4.81	4.87	4.83	0.34	4.44	4.57	4.49	0.64
Low-rated human	4.79	4.50	4.68	0.15	4.02	3.74	3.91	0.28
GPT-3 FS	4.89	4.77	4.84	0.35	3.72	3.67	3.70	0.16
GPT-3 FT-1000	4.50	4.43	4.47	-0.07	2.90	2.78	2.88	-0.34
T5-xxl	4.03	3.68	3.89	-0.76	2.17	2.78	2.25	-0.74

Table 5: Human evaluation scores (out of 5): average on each site (Avg. on ENG/ELL), overall average (Avg.) and average z-score (z) on two criteria, C1 (well-formedness) and C2 (correctness/completeness), as described in the text.

Source	C1		C2	
	First	Last	First	Last
Top-rated human	129	9	104	10
Low-rated human	114	15	68	20
GPT-3 FS	131	5	68	30
GPT-3 FT-1000	97	28	35	62
T5-xxl	71	66	23	90

Table 6: Number of times each system was ranked first (outright or tied) by an annotator, and the number of times it was ranked last (out of 150).

swers containing redundant text. We observed improvements when we used 1000 samples instead of 100 for fine-tuning and hence, fine-tuning on larger data might result in better performance, however, we only experimented with 100 and 1000 samples in this paper due to having limited resources.

Based on Table 3, *T5-xxl* seems to perform similarly to *GPT-3 FT-1000*. However, a small manual evaluation showed otherwise (*GPT-3 FT-1000* answers were slightly better). Furthermore, we observe that the scores for even the best system are very low, but manual evaluations showed that the *GPT-3 FS* generates fairly good answers in many cases. Due to these observations and also given the well-known limitations of automatic metrics for evaluating generation tasks (Kasai et al., 2022; Celikyilmaz et al., 2020; Bhakthavatsalam et al., 2021), we believe conducting human evaluation for deeper analysis is necessary for this task.

In Table 4, we show results for each site to see if one is more challenging than the other. Overall, models perform slightly better on ELL based on automatic metrics—but we see in the next section (Table 5) that there isn’t really a meaningful difference between the sites when humans evaluate the answers.

4.3.2 Human Evaluation

Human evaluators were presented with the question title and body, and then asked to rate 5 answers: a top-rated human-provided answer, a low-rated human-provided answer, and answers generated by 3 of our best models: *GPT-3 FS*, *GPT3 FT-1000*, *T5-xxl*.

They were asked to give ratings (via a slider widget, on a 1–5 integer scale—the higher, the better) for two criteria (C1 & C2):⁸

1. Does the answer look grammatically/structurally like a good answer (ignoring whether it answers the question)?
2. Is the information in this answer a valid response to the question (ignoring formatting/stylistic issues)?

The first criterion aims to get a score for *fluency and coherence* and the second one for *correctness and completeness*.

We collected ratings for a set of 75 questions (375 different answers). Each question with its set of answers was evaluated by at least 2 raters, and then the average score was calculated based on their responses.⁹ We also report the average z-score which is calculated over each annotator’s raw score distribution for each metric, intended to normalize interannotator variation in how the scale is interpreted for each of the two metrics (details in Appendix B).

The results of this study are shown in Table 5. Overall, answers generated by *GPT-3 FS* have a small gap with human answers in both C1 and C2.¹⁰ *GPT-3 FT-1000* comes next, with less accurate answers containing redundant text and hallucinations. The smallest model, *T5-xxl*, ranks last.

Rankings based on human evaluations are available in Table 6. These results are also indicating that model-generated answers are fluent in most cases, but they are not as good as human answers

⁸The survey interface is illustrated in Figure 5 of Appendix C.

⁹Evaluators consisted of 6 English native speakers who are senior NLP researchers and graduate students. The answer source was hidden and the order (5 answers) was randomized. Annotations took about 90 minutes on average.

¹⁰We selected half of the human evaluation samples from very recent posts (June 2021 until January 2023) on Stack Exchange and compared the results with older posts. The models’ scores were comparable or better on the more recent data, so we didn’t see evidence of models having an advantage due to the fact that they may have been trained on some of the data available on the web. For reference, human evaluation scores for recent data were Avg. C1=4.82, Avg. C2=3.83 and for older data, Avg. C1=4.86, Avg. C2=3.61.

Title	Question
Explain the verb tense in "All that glitters is not gold"	I am confused by the maxim All that glitters is not gold . May I know why the word is is used there? In my known according the English grammar glitters is plural, So after plural we add verb+s. As in above example use auxiliary verb so why not are – why is ?
Source	Answer
Top-rated human	Good question, but it's worth pointing out that <i>glitters</i> in this context is a verb, not a noun. The subject of the sentence is <i>all</i> . When <i>all</i> is used as the subject of a sentence, it functions as a pronoun, and it can function as a singular or a plural pronoun. From NOAD <code><blockquote> all (pronoun) [with clause] the only thing (used for emphasis) : all I want is to be left alone. (used to refer to surroundings or a situation in general) everything : all was well \ it was all very strange. </blockquote></code> So: <code><blockquote> All that glitters is not gold </blockquote></code> is roughly equivalent to <code><blockquote> Everything that glitters is not gold </blockquote></code> and the verb tense agrees. but: <code><blockquote> All who eat are not gluttons </blockquote></code> is roughly equivalent to: <code><blockquote> All the people who eat are not gluttons </blockquote></code> so there is no problem there, either.
Low-rated human GPT-3 FS	Grammatically <i>is</i> is used with third person singular and also <i>s</i> form of verb is used. "All" indicates a third person single group. The phrase <i>all that glitters is not gold</i> is an idiom, which means that things that appear to be valuable may not actually be so. It is an old proverb, and the verb tense used in the phrase is an example of the so-called "historic present". The historic present is a literary device used to give a sense of immediacy to a narrative. It is used to describe events that happened in the past, but are being described as if they are happening in the present. In this case, the phrase is used to emphasize the timelessness of the proverb. So, while the verb <i>glitters</i> is indeed plural, the verb <i>is</i> is used in the historic present to give the phrase a sense of immediacy and timelessness.
GPT-3 FT-1000	The maxim is in the present tense, so the verb is also in the present tense: <i>Glitters</i> . The past tense of <i>glitter</i> is <i>glittered</i> , and the past participle is <i>glittered</i> . So if the maxim was in the past tense, it would be: <code><blockquote> All that glittered was not gold. </blockquote></code> The present tense maxim is a generalisation about something that is true in the speaker's experience. The past tense maxim would be a historical statement about what was true in the past.
T5-xxl	This is an interesting question, because it is not immediately obvious what the grammar rules are here. It is not a fixed phrase, and so subject-verb agreement must be based on the context. In this case, the <i>that</i> is a prepositional phrase acting as a complement to the preposition <i>all</i> . It could be rephrased as <code><blockquote> All that glitters is not gold </blockquote></code> It is true that <i>all</i> means "everything that glitters" - but it is equally true that <i>not everything that glitters</i> means "not everything that glitters is gold". It is much the same as <code><blockquote> Everything that glitters is not gold. </blockquote></code> The subject <i>all</i> is singular, so the verb must be singular as well - <i>is</i> . If it was <i>twice</i> plural, then <i>are</i> would be used, but it is not. It is interesting to note that Shakespeare used the same construction in <i>Hamlet</i> : <code><blockquote> Everything that glitters is not gold. </blockquote></code> The subject <i>that</i> is singular, and so the auxiliary verb <i>is</i> agrees with the singular noun.

Table 7: Example post with human and model-generated answers. More examples are available in the Appendix.

when correctness/completeness is considered (*GPT-3 FS* is ranked first or as good as a top-rated human answer in only 45% of cases).

For each criterion and Top-rated human, Low-rated human and *GPT-3 FS*, histograms of the average score of the two annotators are plotted in Figure 3. We can observe that *GPT-3 FS* and Low-rated human have very similar numbers of high-scoring answers (human evaluation scores), but the number of low-scoring human answers drops off gradually as quality decreases, while the distribution is more spread out for *GPT-3 FS*. I.e., the model has some moderately bad answers as well as some good ones, whereas Low-rated human answers cluster more on the upper end of the scale.

C1 (fluency/coherence). All models generated fairly fluent and well-structured answers. We even notice that *GPT-3 FS* scores higher in well-formedness than human answers. We looked at those samples and we believe there are two main reasons for this: 1) Some human answers were very long, containing multiple different quotes from different sources. On average, our evaluators preferred the structure of answers from *GPT-3 FS*, which took the form of a short paragraph addressing the question. 2) Some human answers have a more casual/conversational nature. On the

other hand, *GPT-3 FS* generated more authoritative-sounding, teacher-like answers with complete sentences, which were preferred by our evaluators in some cases.

C2 (correctness/completeness). On average, models are worse on this metric, though sometimes they did produce acceptable answers (perhaps because variants of the question are commonly discussed on the web).

One challenge in this domain is that questions, particularly from ELL, may not be fully fluent. In the *Fluency* example from Table 7 we see that there are some grammatical errors in the question that are unrelated to the topic of the question. In addition, the questioner uses incorrect terminology, mentioning verb tense in the post title even though the question is actually about subject-verb agreement with respect to number (as can be inferred from the examples). While the good human response correctly flags the incorrect terminology and answers the underlying question, *GPT-3* models give irrelevant answers about tense.

Another correctness failure from *GPT-3 FS* can be seen in the following pair involving a *Meaning-to-Form* question:

What is the analog equivalent of 'digitally'?
In electronics, we can solve a problem digitally or analogously? That doesn't sound right, but

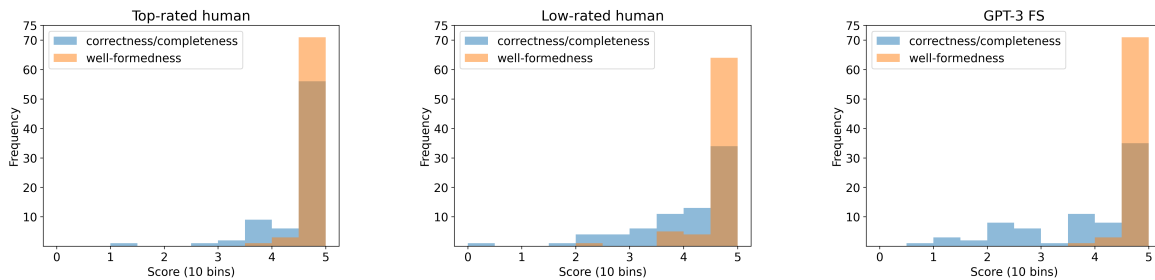


Figure 3: Histograms of the ratings from our human evaluation of answers to 75 questions: the human-authored answer preferred by site users, the human-authored answer least preferred by site users, and our best model (GPT-3 FS). Each value is an average of two annotators’ ratings. (Histograms for GPT-3 FT-1000 and T5-xxl are available in Figure 4 of Appendix C.)

surely there must be a word I can use? **Update:** some definitions: *digitally*: Relating to or being a service that provides information expressed in discrete numerical form. *analoguely*: Relating to or being a service that provides information expressed in continuous, non-discrete form. (**This is my made up definition of a word I made up**). *analogously*: Similar or alike in such a way as to permit the drawing of an analogy. As you can see, the word *analogously* doesn’t mean the opposite of *digitally*, it means something else.

Answer: The analog equivalent of ‘digitally’ is ‘analogously’. Analogously means “in a similar way” or “by comparison”. It is used to describe a situation where two things are similar in some way, but not necessarily opposite. For example, you could say that a digital signal is analogous to an analog signal, because they are both signals, but they are different in terms of how they are represented.

We see that the model begins by ignoring the user’s explanation that *analogously* does not have the desired meaning, and suggests it anyway. The rest of the model’s answer gives a (valid) general definition of *analogously* and clumsily attempts to apply it to the two kinds of signals, which is not responsive to the question. It may be the case that models particularly struggle with questions for which there is no straightforward solution (in this case, no adverb derived from *analogue* has the user’s intended meaning).

4.4 Discussion

4.4.1 Metrics

Our human evaluations found that the high-rated human answers are the best, and *GPT-3 FS* is roughly on par with the low-rated human answer, while other models are worse.

As noted in §4.3.1, we did not find the automatic metrics very informative. We believe this is due to the long and free-form nature of this task: concepts can be explained in different ways and answers can include different examples which make automatic

evaluations very challenging. For example, for the question *Why is this sentence: “Additional nine features were added. . .” incorrect?* a human answer is:

It has to do with the order of the adjectives. For example, consider this sentence: `<blockquote>Happy nine men walk into a bar.</blockquote>` Both *nine* and *happy* are adjectives, but we are really intending *nine* to describe the *happy men*, not *happy* describing the *nine men*. I don’t

Model-generated answers mostly mirror the examples in the question. But answers given by humans mention new examples—which are not in the question (and therefore unlikely to come up in other answers). Thus a good answer may mention superficially off-topic terms metalinguistically, which presents evaluation challenges unlike those of summarization or MT, for example (Reiter, 2018; Celikyilmaz et al., 2020). For QA tasks in metalinguistic domains such as ours, caution is warranted when applying automatic metrics that are unaware of the use/mention distinction. Human evaluation is likely to be more illuminating. We release our full human evaluation data to enable detailed comparisons in future research.

4.4.2 Qualitative Analysis

We examined behavioral patterns for the best model, *GPT-3 FS*, organized by question type, with observations as follows.

Interpretation and Encoding: These are the categories where the model performs the best, especially if the question has a straightforward answer (for example, there aren’t several possible answers based on context). Topics in these categories often center on definitions of vocabulary and idioms. These questions may involve less reasoning to answer, and models may be leveraging content from the online dictionaries and thesauri in their web training data.

Grammatical Analysis: When asked about general concepts and rules of grammar, the model-generated answers are usually acceptable:

-ed and -ing adjectives

What is the difference? <blockquote> I am interesting in mathematics. </blockquote> versus <blockquote> I am interested in mathematics. </blockquote> Murphy intermediate grammar Unit 97 tells me that I cannot use the former. Extended and general explanations will be appreciated.

Answer: The difference between *-ed* and *-ing* adjectives is that the *-ed* adjectives describe a feeling or emotion that is felt by the subject, while the *-ing* adjectives describe a quality of the subject itself. In the example sentence, *interested* is the correct word to use because it describes the feeling of interest that the subject has in mathematics. *Interesting* would be incorrect because it describes the quality of the subject itself, not the feeling of the subject. For example, if you were to say "This book is interesting," you would be describing the book itself, not the feeling you have towards the book. On the other hand, if you were to say "I am interested in this book," you would be describing the feeling you have towards the book.

However, when questions become more specific to examples, sometimes the model struggles:

What does "in" refer to?

<blockquote> A symbiotic relationship is an interaction between two or more species **in** which one species lives in or on another species.

Answer: It refers to the *other* species. The symbiotic species lives **in** the other species.

The example in the question uses the word *in* twice. Rather than addressing the first token, which the questioner has bolded, the answer addresses the second one (denoting the physical relationship between one species and another).

Fluency: The model correctly answers many of the fluency questions. Others require reasoning and understanding of context, and the model struggles to generate a correct answer. For example:

Returning for the *n*th time

Suppose I went to the supermarket three times. Is my third trip considered my "second time returning there" or my "third time returning there?" Thanks. (I know this question sounds silly, but English is my second language...)

Answer: Your third trip to the supermarket would be considered your *third time returning there*. The phrase *returning for the *n*th time* is used to refer to the *n*th time you have returned to a place, not the *n*-1th time.

We also observe that when the correct usage of a case is not mentioned in the question itself, the model sometimes suggests incorrect corrections and wrong explanations:

Why is this usage of "to" incorrect?

I am in communication with a friend who is learn-

ing English. I noticed her using "to" in the wrong context, but I'm unable to explain which rule she is breaking. <blockquote> I can't access **to** Gmail. </blockquote>

Answer: In this sentence, the word *to* is incorrect because it is being used as a preposition. The correct word to use in this context is *for*. <blockquote> I can't access **for** Gmail. </blockquote> The word *to* is used as a preposition when it is followed by a verb, such as *go*, *come*, *run*, etc. For example: <blockquote> I am going **to** the store. </blockquote> However, when *to* is followed by a noun, it is usually incorrect. In this case, the correct word to use is *for*. <blockquote> I need a gift **for** my friend.</blockquote>

4.4.3 Outlook

Overall, we were surprised by the quality of many of the answers from *GPT-3 FS*: many would likely have received high user ratings if submitted as answers on the site. At the same time, the model is not to the point where we would want to trust its answers without human review. We believe that answer confidence estimation—so that users can be shown only the best model-generated answers—is an important direction for using learner QA models in practice (Jiang et al., 2021).

5 Conclusion

We presented ELQA, a dataset containing metalinguistic questions and answers about the English language. We provided analysis and a taxonomy of the data, along with experiments on *free-form answer generation* and investigated the extent to which language models can articulate their generalizations about language. Since many of the questions in ELQA were asked by language learners, it forms a potentially useful and so far untapped resource for educational NLP purposes and metalinguistic question answering. We release the dataset to enable further studies of this task.

Ethics Statement

We have released a processed version of an already public online forum dataset, in a manner consistent with the terms of the license, which require attribution of all posts (§3). The models we have presented are intended only as baselines for future research, not for deployment. Models should be carefully stress-tested for undesirable heuristics/biases before deployment. Systems for the generation task, in particular, would risk misleading language learners with plausible but incorrect answers, so it is important to not deploy a generation

system until it is approximately as reliable as existing non-automated alternatives, and to present the output with caveats. Potential biases reflecting the demographics of authors represented in the training data (in terms of native language, level of English proficiency, etc.) also need to be considered if models are deployed for different target populations.

Limitations

One limitation of our dataset, ELQA, is that the corpus only contains questions in English and about English. However, Stack Exchange has sites with questions about other languages and our main data extraction scripts are general enough that they can be used to create corpora for other sites on Stack Exchange. Of course, language-specific processing steps, quality assurance and analysis must be applied before releasing such data.

Most importantly, the models we have presented here are intended only as baselines for future research, not for deployment. Potential biases reflecting the demographics of authors represented in the training data (in terms of native language, level of English proficiency, etc.) also need to be considered if models are deployed for different target populations. Moreover, many of these types of questions are found on the web, and a lot of the same topics are brought up by many users, so a model’s ability to generate correct answers cannot necessarily be attributed to abstract reasoning.

Acknowledgements

We thank the anonymous reviewers for their insightful comments. We thank Daniel Khashabi for helpful discussions and feedback. This research was supported in part by NSF award IIS-2144881.

References

- Arshad Ahmad, Chong Feng, Shi Ge, and Abdallah Yousif. 2018. [A survey on mining stack overflow: question and answering \(Q&A\) community](#). *Data Technol. Appl.*, 52:190–247.
- Michael L. Anderson, Andrew Fister, Bryant Lee, Luwito Tardia, and Danny Wang. 2004. [On the types and frequency of meta-language in conversation: A preliminary report](#). In *14th Annual Meeting of the Society for Text and Discourse*.
- Sumithra Bhakthavatsalam, Daniel Khashabi, Tushar Khot, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, and Peter Clark. 2021. [Think you have](#)

[solved direct-answer question answering? Try ARCD, the direct-answer AI2 reasoning challenge](#). *arXiv preprint arXiv:2102.03315*.

- Ksenija Bogetic. 2021. [MetaLangCORP: Presenting the first corpus of media metalanguage in Slovene, Croatian and Serbian, and its cross-discipline applicability](#). *Fluminensia*, 33:123–142.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. [Language models are few-shot learners](#). *Advances in neural information processing systems*, 33:1877–1901.
- Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. [The BEA-2019 shared task on grammatical error correction](#). In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.
- Jill Burstein. 2003. The E-rater® scoring engine: Automated essay scoring with natural language processing. In *Automated essay scoring: A cross-disciplinary perspective*, pages 113–121. Lawrence Erlbaum Associates Publishers.
- Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Derru, Mark Cieliebak, and Eneko Agirre. 2020. [DoQA - accessing domain-specific FAQs via conversational QA](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7302–7314, Online. Association for Computational Linguistics.
- Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. [Evaluation of text generation: A survey](#). *ArXiv*, abs/2006.14799.
- Maria Chinkina and Detmar Meurers. 2017. [Question generation for language learning: From ensuring texts are read to supporting learning](#). In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 334–344, Copenhagen, Denmark. Association for Computational Linguistics.
- Robert Dale, Ilya Anisimoff, and George Narroway. 2012. [HOO 2012: A report on the preposition and determiner error correction shared task](#). In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 54–62, Montréal, Canada. Association for Computational Linguistics.
- Cícero dos Santos, Luciano Barbosa, Dasha Bogdanova, and Bianca Zadrozny. 2015. [Learning hybrid representations to retrieve semantically equivalent questions](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on*

- Natural Language Processing (Volume 2: Short Papers)*, pages 694–699, Beijing, China. Association for Computational Linguistics.
- Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. [ELI5: Long form question answering](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.
- Youmna Farag, Helen Yannakoudakis, and Ted Briscoe. 2018. [Neural automated essay scoring and coherence modeling for adversarially crafted input](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 263–271, New Orleans, Louisiana. Association for Computational Linguistics.
- Doris Hoogeveen, Karin M. Verspoor, and Timothy Baldwin. 2015. [CQADupStack: A benchmark data set for community question-answering research](#). In *Proceedings of the 20th Australasian Document Computing Symposium (ADCS)*, ADCS '15, pages 3:1–3:8, New York, NY, USA. ACM.
- Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. [How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering](#). *Transactions of the Association for Computational Linguistics*, 9:962–977.
- Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Lavinia Dunagan, Jacob Morrison, Alexander Fabbri, Yejin Choi, and Noah A. Smith. 2022. [Bidimensional leaderboards: Generate and evaluate language hand in hand](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3540–3557, Seattle, United States. Association for Computational Linguistics.
- Daniel Khashabi, Amos Ng, Tushar Khot, Ashish Sabharwal, Hannaneh Hajishirzi, and Chris Callison-Burch. 2021. [GooAQ: Open question answering with diverse answer types](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 421–433, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Vaibhav Kumar and Alan W Black. 2020. [ClarQ: A large-scale and diverse dataset for clarification question generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7296–7301, Online. Association for Computational Linguistics.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. [Natural questions: A benchmark for question answering research](#). *Transactions of the Association for Computational Linguistics*, 7:452–466.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Edison Marrese-Taylor, Ai Nakajima, Yutaka Matsuo, and Ono Yuichi. 2018. [Learning to automatically generate fill-in-the-blank quizzes](#). In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 152–156, Melbourne, Australia. Association for Computational Linguistics.
- Preslav Nakov, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor. 2017. [SemEval-2017 Task 3: Community question answering](#). In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 27–48, Vancouver, Canada. Association for Computational Linguistics.
- Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. [The CoNLL-2014 shared task on grammatical error correction](#). In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Gustavo Penha, Alexandru Balan, and Claudia Hauff. 2019. [Introducing MANTIS: a novel multi-domain information seeking dialogues dataset](#). *arXiv preprint arXiv:1912.04639*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. [Language models are unsupervised multitask learners](#). *OpenAI blog*, 1(8):9.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Ehud Reiter. 2018. [A structured review of the validity of BLEU](#). *Computational Linguistics*, 44(3):393–401.
- Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, et al. 2022. [Scaling up models and data with t5x and seqio](#). *arXiv preprint arXiv:2203.17189*.

- Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. [How much knowledge can you pack into the parameters of a language model?](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.
- Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. [QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension.](#) *ACM Comput. Surv.*, 55(10).
- Keisuke Sakaguchi, Yuki Arase, and Mamoru Komachi. 2013. [Discriminative approach to fill-in-the-blank quiz generation for language learners.](#) In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–242, Sofia, Bulgaria. Association for Computational Linguistics.
- Charlotte Taylor. 2015. [Beyond sarcasm: The metalanguage and structures of mock politeness.](#) *Journal of Pragmatics*, 87:127–141.
- Kristiyan Vachev, Momchil Hardalov, Georgi Karadzhov, Georgi Georgiev, Ivan Koychev, and Preslav Nakov. 2021. [Generating answer candidates for quizzes and answer-aware question generators.](#) In *Proceedings of the Student Research Workshop Associated with RANLP 2021*, pages 203–209, Online. INCOMA Ltd.
- Yu Wang, Yuelin Wang, Kai Dang, Jie Liu, and Zhuo Liu. 2021. [A comprehensive survey of grammatical error correction.](#) *ACM Trans. Intell. Syst. Technol.*, 12(5).
- Shomir Wilson. 2010. [Distinguishing use and mention in natural language.](#) In *Proceedings of the NAACL HLT 2010 Student Research Workshop*, pages 29–33, Los Angeles, CA. Association for Computational Linguistics.
- Shomir Wilson. 2011. [In search of the use-mention distinction and its impact on language processing tasks.](#) *IJCLA*, 2(1-2):139–154.
- Shomir Wilson. 2012. [The creation of a corpus of English metalanguage.](#) In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 638–646, Jeju Island, Korea. Association for Computational Linguistics.
- Shomir Wilson. 2013. [Toward automatic processing of English metalanguage.](#) In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 760–766, Nagoya, Japan. Asian Federation of Natural Language Processing.
- Shomir Wilson. 2017. [A bridge from the use-mention distinction to natural language processing.](#) In Paul Saka and Michael Johnson, editors, *The Semantics and Pragmatics of Quotation*, pages 79–96. Springer International Publishing, Cham.
- Yuan Yao, Hanghang Tong, Tao Xie, Leman Akoglu, Feng Xu, and Jian Lu. 2013. [Want a good answer? Ask a good question first!](#) *arXiv preprint arXiv:1311.6876*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [BERTScore: Evaluating text generation with BERT.](#) In *International Conference on Learning Representations*.

A Data Credits

The Stack Exchange license requires that any Internet use of the content should include a hyperlink directly to the original question and the profile of the authors. Below are URLs for all the examples used in this paper. The post URL incorporates the post title.

- <https://ell.stackexchange.com/questions/12/dates-and-times-on-in-at> (Q by bytbuster, A by waiwai933)
- <https://ell.stackexchange.com/questions/146633/on-my-own-way-vs-in-my-own-way> (Q by bavyan-yaldo)
- <https://ell.stackexchange.com/questions/19684/wording-what-get-by-means-in-this-context> (Q by nima)
- <https://english.stackexchange.com/questions/74896/grammatically-correct-synonym-for-level-of-catastrophiness?> (Q by solvingPuzzles)
- <https://english.stackexchange.com/questions/134352/should-i-modify-a-gerund-using-an-adjective-or-an-adverb> (Q by worawit-tepsan)
- <https://english.stackexchange.com/questions/222567/what-is-the-etymology-of-physician> (Q by casvaart)
- <https://ell.stackexchange.com/questions/185516/did-vs-have-done> (Q by learner)
- <https://english.stackexchange.com/questions/162824/what-is-the-analog-equivalent-of-digitally> (Q by rocketmagnet, first A by AllisonAshley, second A by Hot Licks)
- <https://ell.stackexchange.com/questions/13749/explain-the-verb-tense-in-all-that-glitters-is-not-gold> (Q by Chinmay235, first A by J.R., second A by sajad)
- <https://english.stackexchange.com/questions/162824/what-is-the-analog-equivalent-of-digitally> (Q by Rocketmagnet)
- <https://english.stackexchange.com/questions/203518/why-is-this-sentence-additional-nine-features-were-added-incorrect> (Q by user95069), A by Nick2253
- <https://english.stackexchange.com/questions/49384/ed-and-ing-adjectives> (Q by itun)
- <https://ell.stackexchange.com/questions/87725/what-does-in-refer-to> (Q by Anfi)
- <https://english.stackexchange.com/questions/102996/returning-for-the-nth-time> (Q by AlicornTwilightisaTroll)
- <https://english.stackexchange.com/questions/55331/why-is-this-usage-of-to-incorrect> (Q by Ademos)
- <https://ell.stackexchange.com/questions/87725/what-does-in-refer-to> (Q by Anfi)
- <https://ell.stackexchange.com/questions/322637/he-is-more-than-a-friend-is> (Q by Loviii, first A by MarcInManhattan, second A by Kirt)
- <https://english.stackexchange.com/questions/258060/verb-for-doing-something-unknowingly> (Q by Daniel Bramhall, first A by chasly - supports Monica, second A by talrnu)
- <https://ell.stackexchange.com/questions/322580/know-someone-in-detail> (Q by Simo Ita)

B On our use of z-scores

In our human evaluation, raters were presented with a question and five candidate answers and asked to

rate each on a scale from 1 to 5 for each of our two criteria (C1 and C2).

Our main goal is to compare the quality of the answers across 5 conditions (3 systems, 2 posts from the site). Raters may have different interpretations of the absolute scales—for example, some raters could be more generous than others overall in terms of the numerical rating, even if they agree on the ranking of systems.

There are several possible ways to factor out this bias. One way is to compute standard scores, a.k.a. z-scores, for each annotator’s distribution of responses on each criterion. Consider C1: from the ratings of an annotator a we have the empirical distribution

$$P_a^{C1}(y_{i,a}^{C1} | x_i)$$

where i indexes the items (answers, of which multiple ones may belong to the same question), and likewise for C2. For each of these distributions we fit a normal distribution by computing mean and standard deviation. For an absolute rating $y_{i,a}^{C1}$, its z-score $z_{i,a}^{C1}$ is its number of standard deviations above the mean rating for that annotator on that metric (a negative z-score indicates it is below the mean). Averaging the z-scores for a particular condition, we can see whether annotators tended to rate outputs in that condition with higher or lower scores than the other outputs they saw in the sample. Note that the z-score computation ignores the grouping of answers from different conditions into questions, so it is not directly measuring annotators’ rankings of candidate answers to a particular question.

C Further Details

Parameter	Value
Batch Size	8
Max. Gradient Updates	75k
Max. Input Length	512
Max. Output Length	512

Table 8: T5 hyperparameters used for the Answer Generation task

	ROUGE-1	ROUGE-2	ROUGE-L	BLEU Score	BERTScore
Set-1	0.303	0.084	0.201	0.092	0.859
Set-2	0.296	0.079	0.192	0.092	0.854
Set-3	0.286	0.071	0.193	0.052	0.856
Set-4	0.273	0.067	0.188	0.039	0.857
Set-5	0.289	0.074	0.193	0.051	0.856

Table 9: GPT-3 few-shot scores on a subset of dev when using different sets of in-context examples

Title	Question
"He is more than a friend is"	(1a) <i>He is taller than his friend.</i> — correct (1b) <i>He is taller than his friend is.</i> — as I know, (1b) is also correct, although less common than (1a) (2a) <i>He is more than a friend.</i> — correct (2b) <i>He is more than a friend is.</i> — by analogy with (1b), (2b) must be correct too. Is it so? If not, then why is (2b) not correct, whereas (1b) that is analogous to (2b) is correct?
Source	Answer
Top-rated human	This is a bit tricky, because "than" can be a conjunction or a preposition, and not everyone agrees on when it is each. I'll present one interpretation that I believe is pretty standard. In 1a and 1b, "than" is best understood as a conjunction. It compares two actions ("he is taller" and "his friend is tall"). That is why we can insert "is" at the end of 1b. In 2a and 2b, "than" is best understood as a preposition. The prepositional phrase "than a friend" modifies "more". It does <i>not</i> compare two actions ("he is more" and "a friend is [adjective]"). That is why we can not insert "is" at the end of 2b. To summarize: In your sentences 1a and 1b, "than" is a conjunction. In 2a and 2b, it is a preposition. Thus, the situations are not analogous.
Low-rated human	(1a) He is taller than his friend. — correct, but incomplete, and natural (1b) He is taller than his friend is. — correct, but incomplete, and natural (1c) He is taller than his friend is tall. - correct and complete, but awkward (2a) He is more than a friend. — correct, but incomplete, and natural (2b) He is more than a friend is. - correct, but incomplete, and awkward (2c) He is more than a friend to me. - correct, but incomplete, and natural (2d) He is more than a friend is to me. - correct, but incomplete, and awkward (2e) He is more to me than a friend is to me - correct, complete, and awkward The complete statement of meaning is grammatically correct but may be more, or longer, than people naturally say. Typically words that can be implied are eliminated for more natural speech. Some eliminations of implied words will lead to constructions that are grammatically correct and natural, while other eliminations of implied words can lead to constructions that are grammatically correct but which sound awkward. The awkwardness of an expression has more to do with common usage than it does with rules of grammar.
GPT-3 FS	No, (2b) is not correct. The reason is that (2a) is a comparison between two things, while (2b) is not. In (2a), the comparison is between <i>he</i> and <i>a friend</i> , while in (2b) the comparison is between <i>he</i> and <i>his friend is</i> . This is not a valid comparison, so (2b) is not correct.
GPT-3 FT-1000	The difference between (1a) and (1b) is that (1a) is a statement about <i>the friend</i> , whereas (1b) is a statement about <i>the friendship</i> . Consider the following two sentences: <blockquote> John is taller than his friend Bob. </blockquote> vs. <blockquote> John is taller than Bob is. </blockquote> The first sentence is a statement about John and Bob. The second sentence is a statement about John and the friendship between John and Bob. In the first sentence, the friend is the subject of the sentence. In the second sentence, the friend is the object of the sentence. In your examples, (1a) and (2a) are statements about the friend, whereas (1b) and (2b) are statements about the friendship.
T5-xxl	Answering your question, I can say that, no, (2b) is not correct. (2a) "He is taller than his friend" is idiomatic and natural, and the simple present tense is used here to describe an action or state. (2b) "He is more than a friend" is not idiomatic and does not mean the same thing.
Title	Question
Verb for doing something unknowingly	I cannot think of an effective verb that would suggest someone is doing something unknowingly yet doing it nonetheless - almost like acquiescing. I have thought of 'sleepwalking' however there must be something better. For example, VERB into a materialistic society.
Source	Answer
Top-rated human	They are drifting into a materialistic society. They are sliding into a materialistic society.
Low-rated human	<i>Evolve</i> fits your needs and in particular your example pretty closely. Evolution is a natural process, implying the lack of intention you're looking for. It also fits your example as it's an act of change and development. Similarly, you might consider <i>mutate</i> , another natural occurrence with greater emphasis on randomness and less emphasis on a direction of development. More generally, you could say my answer is: use a context-specific verb which emphasizes nature (i.e. the lack of human intention) as the driving force behind the action.
GPT-3 FS	The verb <i>drift</i> could be used in this context. It suggests a lack of conscious direction or control, and implies that the person is being carried along by external forces. <blockquote> He <i>drifted</i> into a materialistic society.</blockquote>
GPT-3 FT-1000	I think <i>bemused</i> would work well in this context. It means to confuse or bewilder, but can also be used to describe someone who is behaving in a confused way, or who is reacting to something confusing. So, "Jane was bemused by the materialistic society she found herself in." or "Jane was bemused by her own behaviour as she bought yet another pair of designer shoes.
T5-xxl	How about <i>slumbering</i> into a materialistic society?

Table 10: Two example posts with human and model-generated answers.

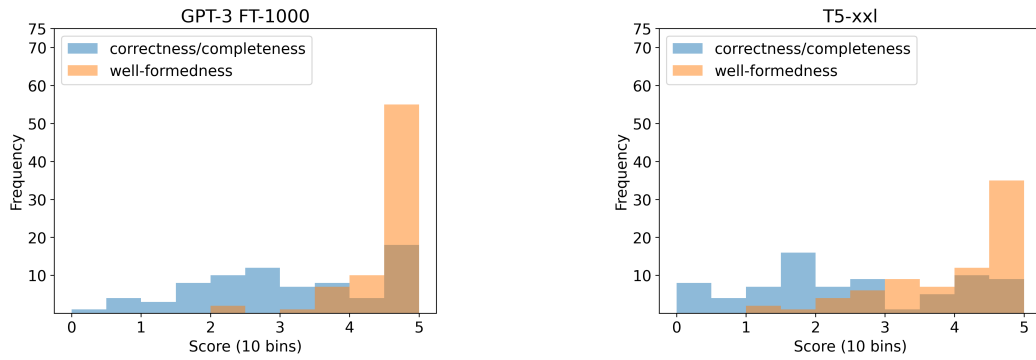


Figure 4: Histograms of the average score of each two annotators from our human evaluation for answers generated by GPT-3 FT-1000 and T5-xxl. The other answers evaluated are represented in Figure 3.

Q6:
Title: Know someone in detail
Body:
 Is it possible to say **know someone in detail** in English?

In my opinion, if you have only one best friend, you can develop a deep friendship more easily. On the other hand, if you have a lot of friends, you don't have the opportunity **to know them in detail**.

If not, what would you suggest instead?

Answers:

Know can be used to mean *become familiar with* . It can also mean *become close to* .

I don't know my new boss in detail. He's a complete stranger.
 I don't know my old boss in detail. He's a complete stranger.

The meanings are quite different. To *know* something is to understand it, to be familiar with it, to have it be a part of your life. To *know* something is to own it, to be responsible for it, to be intimately familiar with it.

0 5

Does the answer look grammatically/structurally like a good answer (ignoring whether it answers the question)

4

Is the information in this answer a valid response to the question (ignoring formatting/stylistic issues)

1

Figure 5: Screenshot from the survey we designed for human evaluation on the Qualtrics platform. Five answers were shown for each question as described in §4.

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
7
- A2. Did you discuss any potential risks of your work?
6
- A3. Do the abstract and introduction summarize the paper’s main claims?
1
- A4. Have you used AI writing assistants when working on this paper?
Left blank.

B Did you use or create scientific artifacts?

3

- B1. Did you cite the creators of artifacts you used?
3, *Appendix*
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
3
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
3
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
3
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3.1, 4.1

C Did you run computational experiments?

4.2

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4.2, *appendix*

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4.2
- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4
- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
4
- D** **Did you use human annotators (e.g., crowdworkers) or research with human participants?**
4
- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
4, *appendix*
- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
4
- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
3,4
- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.
- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
4 (*footnote*)