

CLEAR : UNDERSTANDING THE REASONING CAPABILITIES OF LARGE LANGUAGE MODELS

Anonymous authors

Paper under double-blind review

ABSTRACT

Despite significant progress, accurately assessing the reasoning capabilities of Large Language Models (LLMs) remains both a challenging and divisive subject. Many existing benchmarks either suffer leakage, or reflect patterns in the training data, leading to ambiguous results. We present CLEAR (Conlang Logic Evaluation And Reasoning), a novel benchmark designed to test the reasoning and problem solving capabilities of LLMs in new environments. CLEAR uses Conlangs (Constructed Languages) for few-shot translation tasks, which require some linguistic knowledge to solve, but primarily the ability to make new patterns from tokens in unfamiliar contexts using logical operations. These conlangs represent a unique challenge, as while translation examples are plentiful, these conlangs each have a unique combination of rules, are self contained, and are absent in the training corpus. We present an evaluation of current frontier models over multiple metrics as a baseline for future research. We will be releasing CLEAR as a public benchmark to drive progress towards AI systems more capable of general reasoning.

1 INTRODUCTION

Evaluating Large Language Models (LLMs) is a challenging task. Evaluating the abilities of LLMs to reason, is both challenging and divisive. With the use of LLMs rapidly filtering into many areas of life, from code completion (Rozière et al.) to designing proteins (Li et al.), it has never been more important to evaluate these models robustly.

We feel that insight into the true reasoning capabilities of LLMs is lacking, largely due to the complexity of measuring this performance. As this is such a nascent area of study there are many complimentary measures of reasoning, which can broadly be categorised into two domains. One is to take the contents of an advanced subject and correctly draw the right conclusions to complex problems (Rein et al.). This is crucial in areas such as law, scientific analysis, and mathematics to name a few, and will be familiar to anyone who has used an LLM as an education tool.

However, while this level of accomplishment across a field is already arguably beyond that of many humans, it does not measure the ability to adapt to new information, which for people is second instinct. This is the second kind of intelligence to evaluate. How well can the model adapt to new information, leverage what is already known, and yield correct reasoning. This second intelligence is the key component required for the kinds of scientific discovery and path towards a more general system, and we argue is under evaluated in current benchmarks. Unfortunately it is largely under evaluated because it is a very challenging problem to solve, how do you find a task that challenges the LLM that is not represented in the training data? One such attempt is the Abstract Reasoning Corpus (Chollet) which provides LLMs with geometrical problems to solve, which are visually intuitive to a human but challenging to the LLM, and crucially absent from the training data. We build on this philosophy with the proposed CLEAR benchmark.

Here, in our attempt to evaluate the reasoning abilities of LLMs, we focus on one very specific task, learning a new language. We provide the model with an example of an unseen constructed language (Conlang) with translation sentences in English and some preliminary information describing how the language works, then ask for translations to example sentences. A cartoon of this setup is shown in Fig. 1. Such constructed languages have been predominantly been found in language learning enthusiast circles as toy languages but can exhibit all the features of real natural languages, while also having the advantage of not being present in the training set. The simplicity of this task will

054
055
056
057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095
096
097
098
099
100
101
102
103
104
105
106
107

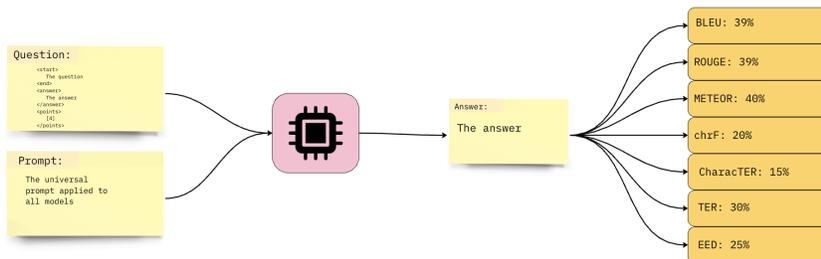


Figure 1: The translation workflow is shown here, where the question and universal prompt and provided to the LLM, and the answer evaluated with the suite of scoring mechanisms.

feel intuitive to almost everyone, when provided with only a few example translations patterns and repeated words start to become apparent, regardless or prior exposure to the language.

To ‘learn’ the rules of such a language is relatively simple. The translation tasks require a basic understanding of the grammatical rules and patterns with which languages are generally manipulated, and an understanding of simple logical rules. (We will revisit this concept in Sec.3.1.) Both are areas we expect modern LLMs to be proficient in considering knowledge based evaluations - this is something we aim to provide additional clarity on here. As such when the model sees an English sentence paired with a Conlang sentence we expect one sentence to be well tokenized and have good prior embeddings, and the other to be much more random. To extract a meaningful translation the model must be able to update the prior embeddings for the Conlang tokens, and make new mappings. An advantage of this task is that translation tasks are prevalent in the training data, so the pattern should be well understood, the complexity comes almost entirely from having to impose meaning onto a set of poorly initialised tokens.

We believe this poses a promising route to evaluate the in context learning capabilities of LLMs. These translation tasks show the extent to which the models can (1) use simple logical rules, to (2) update the meaning of tokens based on new information. This offers insight into the in-context learning abilities of the model, and to the reasoning ability of models to update beliefs based on new information.

Authors Key Contributions: In this paper we present three key contributions:

- Firstly we proposed a new dataset CLEAR to benchmark the ability of LLMs to reason through the medium of a translation task on new unseen languages.
- Additionally we provide initial benchmark scores for a range of current LLMs.
- Finally we also provide some analysis on the successes and limitations of current models.

2 RELATED WORK

Few-Shot Prompting of Large Language Models Vaswani et al. showed the promise of LLMs for machine translation tasks, and Brown et al. (2020) introducing GPT-3 showed that LLMs conditioned with several previously unseen examples is able to adapt to novel tasks without updating model parameters, in a method known as few-shot prompting.

Posing all tasks as sequence to sequence problems as in Chowdhery et al. (2022) and Raffel et al. (2023) makes the task of injecting few shot prompts simple to apply to a wide range of tasks, including translation. Multilingual models have been shown to exhibit the same behaviour (Lin et al., 2022) even in low resource cases, due to the ability to utilize high resource language pairs.

All current frontier models are trained on a diverse mix of natural languages (Touvron et al. Dubey et al. OpenAI et al.), which has yielded improved translation ability. Although as was found in Workshop et al. the multilingual ability is strongly correlated with data quality, particularly in low

108 resource languages. However, despite strong progress, translation of completely unseen languages
109 requires either reasoning capability, or significant utilisation of other translation tasks in the training
110 corpus.

111 **Logical Reasoning of Large Language Models** Recent work has shown signs that LLMs are ca-
112 pable of performing tasks that require logical reasoning and problem solving, despite largely being
113 trained on next-token-prediction, many benchmarks show exceptional exam performance in fields
114 such as law (OpenAI et al.), PhD level scientific domains (OpenAI, 2024), and mathematics (Zhang
115 et al.). The source of this phenomena is an area of open debate, as either representative of an internal
116 world view, or repeating patterns from the training data. Many examples claim to prove or disprove
117 this using, for example, classic riddles and then twisting the setting subtly.

118 More concretely, Wei et al. proposed the *Chain of Thought* method, where the model is asked to
119 generated additional outputs during generation akin to a stream of thought, and has generally been
120 found to improve performance on tasks requiring multi step reasoning. Kojima et al. demonstrated
121 similar results using additional prompts like *"think step by step"*. It is argued that these results
122 show the inherent reasoning in LLMs. Taking this approach, further work such as Zelikman et al.
123 (2022) and Zelikman et al. (2024) starts to train a iterative loop to improve model outputs, it has
124 been speculated that this kind of approach is behind the latest developments in OpenAI (2024).
125 The mechanism for this in context learning is still unknown with arguments proposing effective in
126 context gradient descent (Shen et al., 2024), however the true mechanism remains unknown.

127 **Logical Reasoning** There are many benchmarks quantifying aspects of the reasoning abilities of
128 these models. Notable attempts include Srivastava et al. where a collection of tasks that require
129 multiple steps of reasoning to complete has been organised. Other benchmarks are more specific,
130 for example consists of questions on a short text where indirect relationships must be extracted, Liu
131 et al. relies on a multiple choice approach, this is nice as it allows for subtle differences to be probed,
132 and Tafjord et al. goes further requiring the model to produce from a set of facts the logical steps
133 that justify the reasoning. Law (OpenAI et al.) and scientific (OpenAI, 2024) exam questions partly
134 address this reasoning, but are hard to isolate from genuine facts present in the training corpus. Even
135 strong cutting edge benchmarks such as GPQA(Rein et al.) have limitations, and are more focused
136 on the ability to solve complex domain specific questions, which relies on specific knowledge and is
137 subject to leakage. The different attempt by Chollet formulates geometric problems that are outside
138 any training corpus as a hard evaluation problem.

139 Similarly to the step-by-step approach, many examples exist where a scratchpad is provided to the
140 LLM to output a set of intermediate tokens in an attempt to allow the model to guide towards a local
141 minima (Nye et al.). However, there is increasing evidence (Ye & DurrettTurpin et al.) that while
142 the use of intermediate tokens can help reach a better output, the intermediate explanation or chain
143 of thought is subject to just as many "hallucinations" as other outputs. Recent evaluations find the
144 latest generation of LLMs still fail to plan multiple steps sufficiently (Valmeekam et al.).

145 **Translation Evaluation Metrics** Machine Translation (MT) requires automated metrics to operate
146 at scale and to provide consistent scoring mechanisms. Traditional metrics such as BLEU (Papineni
147 et al., 2002), METEOR (Banerjee & Lavie, 2005), ROUGE (Lin, 2004) have survived the test of
148 time largely due to their simplicity. BLEU is arguably the most widespread as both the simplest,
149 performing n-gram checks, and the most limited, not providing any flexibility in translation out-
150 put through word stems or synonyms. As there are often many equally valid possible translations,
151 alternative methods attempt to accommodate this. BERTScore (Zhang et al., 2020) uses a BERT
152 (Devlin et al., 2019) model to compute similarity scores between candidate and target sentences us-
153 ing contextual embedding, allowing for more sophisticated translations. COMET (Rei et al., 2020)
154 takes this approach further by using a dedicated neural network to judge the translation. However,
155 such methods introduce their own limitations, and may perform unpredictably on languages not
156 represented in their training corpus.

156 Other metrics such as chrF (Popović, 2015) and chrF++ (Popović, 2017) take the n-gram approach
157 to characters, offering more granularity for translations missing subtle features in the output. Addi-
158 tionally metrics such as YiSI (Lo, 2019) attempt to align outputs closer to human preference.

159 There are also edit based methods that evaluate the quality of the translation through the number
160 of edits required to convert the candidate sentence into the reference. Metrics such as TER (Panja
161 & Naskar, 2018) count the number of insertions, deletions, shifts, etc required to match the target

162 sentence. EED (Stanchev et al., 2019) extends this to better capture human preference, and CDER
 163 (Leusch et al.) and CharacTER (Wang et al., 2016) operate at the character level making them
 164 suitable to languages with rich morphology, and possibly robust to tokenization issues in LLMs.
 165

167 3 DATASET CREATION

169 3.1 TRANSLATION EXAM SCRIPTS

172 The exam scripts were collated from online resources used primarily as entrance examinations for
 173 university level language degrees, however they are also popular within language learning commu-
 174 nities online. Students applying to study Modern Foreign Languages at Universities such as Oxford
 175 University must take an entrance exam. A part of that exam is the Language Aptitude Test, which
 176 is designed to evaluate students on their ability to parse and understand a new unseen language that
 177 doesn't rely on real world knowledge.

178 The motivation is to standardize the test such that prior knowledge of the language doesn't provide
 179 some applicants an advantage over others, and the results should correlate with likely ability to study
 180 new languages. We propose to take this motivation and apply it to LLM evaluation.

181 Each exam would consist of a preamble describing the translation task, proving some details to help
 182 understand how to approach the conlang (i.e. is the language independent of word order, are the
 183 nouns gendered, etc.) then split into three or four sections. The sections then each provide a set of
 184 translation sentences, with increasing complexity as the exam progresses. Each section then has a
 185 number of questions translating from English to the Conlang, and from the Conlang to English. All
 186 translation sentences are possible to derive provided the information from the previous sections has
 187 been understood. This format maps well onto a few-shot prompt for an LLM.
 188

189 3.2 METHODOLOGY

192 The dataset was constructed by scraping the publicly available exam scripts and their answers, and
 193 parsing them into individual questions as prompts for the language models. Questions for the each
 194 section all share the same preamble and translation examples, but have a different target sentence to
 195 translate. For example where section (a) might have 10 example sentences and 4 translation tasks,
 196 we repeat the stem and examples as the prompt for each of the 4 translation tasks. While there is
 197 some duplication of analysis here, this tests the ability of the model to extract the relevant patterns
 198 in different views, and the preliminary sections are required to solve the more complex later parts of
 199 the exam. As the exams are intended to be conducted as a complete exercise each question builds
 200 on knowledge derived for previous answers. Therefore for each subsequent section we append the
 201 new translation examples to the previous question stem and previous translations. This means the
 202 model has access to the example translations for (a) + (b) when answering (b), and (a) + (b) + (c)
 203 when answering (c). etc.

204 A simplified example of a question is given in Fig. 2. Here we highlight the patterns we might expect
 205 to extract as a human attempting to translate this task. We might start by trying to understand the
 206 word order, then find some patterns that allow us to start matching candidate word stems. Then start
 207 thinking about the endings / grammatical changes to the words based on their contexts.
 208

209 3.3 DATASET STATISTICS AND CHARACTERISTICS

212 The dataset consists of 140 questions, from 12 exam scripts. There are 41 questions worth 3 points,
 213 55 questions worth 4 points, 31 questions worth 5 points, and 13 questions worth 6 points. The
 214 maximum exam score is therefore 576 points. The complexity of the translation increases with the
 215 increase in points, and we reflect this in our 'Exam' metrics during evaluation. Of the 140 questions
 72 are ENG→CON and 68 are CON→ENG .

216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231

Example Question: (Fub) Word order is not fixed, and there is no distinction between a / the.

huufts hohub red = The teachers taught the pupils.
hip pik hohub = A teacher provided homework.
rored pik daawl = The pupils had homework.
rored pik liik = The pupils liked the homework.
tok daaw rored = A pupil had a dog.
totok liiks red = The dog liked the pupils.
tok rored huuf = The pupil taught the dog.
paat pik totok = The dog ate the homework.

Translate the following sentence(s):
liik hohub tok
Answer =
totok hub liiks
Answer =

Reasoning:

(1) liik / liiks = liked, looks like a plural vs singular
(2) hohub = Teacher / Teachers - A / The no distinction, and the plural seems to be set by the verb
(3) tok / totok = dog - subject vs object of the sentence

Therefore:
liik hohub tok = the teachers liked the dog

The teachers (from hohub, and plural from liik) liked (from liik) the dog (from tok, subject)

totok hub liiks = The dogs liked the teachers

The dogs (subject, from totok vs tok, plural from liiks) liked (from liiks, this also gives plural to subject) the teachers (from hub, object from hub, vs subject hohub)

232 Figure 2: An example question from the dataset on the left with a (simplified) description of the
233 language specific features. Two translation sentences are provided, but note in the real dataset only
234 one is present at a time. On the right we show a simple way of reasoning towards the correct answer.
235 Words are colour-coded to show which words are linked across the translation, and a more detailed
236 description of how the translation is reached is given. This should give an indication of how a human
237 might approach the task, we can see how it lends itself to a few-shot format iteratively building the
238 information needed to solve the questions.

241 4 METHODOLOGY

244 4.1 MODELS EVALUATED

246 We have attempted to evaluate a broad a range of LLMs at a variety of model scales, ranging from
247 the latest frontier models to small optimised LLMs. While any selection will inevitably leave some
248 models out, this choice was made to provide representative coverage for typical use cases across a
249 range of scales. The models evaluated fall into three categories:

250 **Closed Frontier Models:** These are models such as the latest GPT4o (noa), Claude (Anthropic,
251 2024), Gemini (Deepmind, 2024), and Mistral (Mistral, 2024). These models are broadly considered
252 at be leading in terms of their capacities across a variety of benchmarks. These models, however,
253 are opaque in terms of their exact architecture, prompt construction, and size. These models are all
254 accessed via API calls.

255 Models: *GPT-4o*, *GPT-4o-Mini*, *GPT-3.5 Turbo*, *GPT-4 Turbo*, *Claude Opus*, *Claude Sonnet*,
256 *Claude Sonnet 3.5*, *Gemini-1.5 Pro v1/2*, *Gemini-1.5 Flash v1/v2*, *Mistral Large*, *Mistral Small*

258 **Optimised Open Source Models:** We also evaluate a number of open source models. This is made
259 possible on moderate hardware using OSS efforts such as `ollama` (Collaboration, 2024). These
260 models are typically quantised, and efficiently reduced in size so they can be used on most modern
261 laptops. While these optimisations will incur some small reduction in performance, it was decided
262 this was a suitable trade off for the evaluation as this represents a common use case for LLM use in
263 the real world.

264 Models: *LLama 2 70bn* (Touvron et al.), *Gemma 2 - 7bn*, *Gemma 2 2 bn* (Team et al.)

265 **OpenAI: GPT-4 o1 Preview** Additionally we also studied the preview of the GPT-4-o1 (OpenAI,
266 2024) model from OpenAI. This model however was not available to us via the API, therefore we
267 used the ChatGPT application to provide prompts one at a time. Each question was provided in a
268 new chat so no memory persisted between questions. At this time, due to rate limits we have only
269 been able to evaluate 80/140 of the questions, as such results will be provided with an asterisk until
this is complete.

4.2 EVALUATION METRICS

In our metrics we are looking for a number of properties, which are slightly different to usual translation cases. Typical translation metrics have moved away from exacting n-gram matches (Papineni et al., 2002) to allow for more flexible and natural translation (Zhang et al., 2020). While this is a significant improvement in natural language translation, we want to quantify the translation using harsh scores in order to evaluate the logical reasoning behind the translation.

The Conlang translations do not require creative solving, the vocabulary should be inferred from that used in the question stem, and the subtle variations to answers are often exactly the small grammatical errors that we are trying to evaluate the models on. To that end we have chosen the following metrics.

BLEU: This is our main metric, BLEU (Papineni et al., 2002) is a longstanding popular metric in machine translation due to its simplicity and interpretability. Additionally it has no dependence on the language, and the strong penalty for mistakes is preferable in the context of looking for exact translations of very tightly controlled sentences.

The BLEU metric computes p_n (modified) n-gram precision scores for lengths up to N , weighted by w_n , and is given by

$$\text{BLEU} = \begin{cases} \exp\left(\sum_{i=0}^N w_n \log p_n\right) & \text{if } c > r \\ \exp\left(1 - \frac{r}{c} + \sum_{i=0}^N w_n \log p_n\right) & \text{if } c \leq r. \end{cases} \quad (1)$$

This score is modified by the brevity term which penalises the output sentence of length c for being shorter than the target sentence of length r . This gives us $\text{BLEU} \in [0, 1]$. As BLEU penalises any mismatching words harshly, we enforce a strict measure on difference between words like ‘horse’ vs ‘horses’ or ‘run’ vs ‘ran’ as while the meaning may be semantically similar the error implies the underlying grammar has not been understood.

ROUGE: (Recall-Oriented Understudy for Gisting Evaluation) This is another n-gram matching metric, where in contrast to BLEU the main focus is recall. Recall is important in evaluating translations, as it would be possible to get a very high precision score for the produced translation, but to miss a significant fraction of the translation. BLEU takes this into account with the brevity penalty, but is an imprecise measure of how much the translation covers the target (Lin, 2004). This makes it less focused on the precision of the translation, but provides a different angle to to evaluate the outputs.

METEOR: The principle is to combine precision (P) and recall (R) evaluation into a single metric, uni-grams are matched between the candidate and the reference sentences and the number of matches computed (Banerjee & Lavie, 2005).

$$\text{METEOR} = \frac{10 \cdot P \cdot R}{9 \cdot P + R} \left(1 - 0.5 \frac{\#chunks}{\#matches}\right) \quad (2)$$

The METEOR score has a different trade off to the BLEU score, here semantic variations are allowed through the use of word stems and paraphrases. There is an obvious drawback to metrics like METEOR in this case, as the more sophisticated stages of the algorithm rely on finding matches to the stems of words, or synonyms. This is obviously not possible for Conlangs, and we will have to fall back to the basic METEOR implementation for $\text{ENG} \rightarrow \text{CON}$ translations.

chrF: (Character n-gram F score) follows a similar philosophy to BLEU, but computes the n-grams on the character level instead. Typically n-grams of up to 6 are chosen and the harmonic mean of precision and recall is computed. It’s generally found that for morphologically rich languages, a low chrF score correlates with human preferences over metrics such as BLEU which focus on the word level ?. The metric is computed as:

$$\text{chrF} = (1 + \beta^2) \cdot (P \cdot R) / (\beta^2 * P + R), \quad (3)$$

where β controls the weighting between precision and recall. The drawbacks with this metric in common use cases are advantageous in evaluating the translations in CLEAR, the metric is sensitive to very small changes which while they may not fundamentally harm understanding, they show

where the models have not parsed the logic correctly, and the lack of semantic variation keeps the metrics clean for the Conlangs.

EED: (Edit Evaluation Distance) A similar concept to chrF, where the score reflects the number of edits required to transform one sentence into another. Edits include insertions, substitutions, deletions, etc. and a lower score is better (Stanchev et al., 2019).

TER: (Translation Edit Rate) This metric takes the same concept as EED, but looks at the word level for transformations, in a method that is designed to mirror more closely the process a human translator might take (Snover et al.).

CharacTER: (Character-level Translation Edit Rate) A culmination of the above metrics, which applies the translation edit rate approach to the character level, but with awareness of the word boundaries through the addition of special characters between words (Wang et al., 2016). This results in a more sensitivity when compared to word level metrics, and better provides partial credit for translations.

Exam Setting: We also introduce a weighted version of some of the metrics to account for the question complexity and output sentence length. Later questions in the exams can be worth up to 6 points, compared to the 3 points per question for the early parts. Therefore we weight the output of some of the metrics (those whose outputs naturally scale between 0 and 1) by the points associated with that question, and compute a score over the entire dataset. As detailed in Sec.3.3 there is a possible total of 576 points, but for simplicity we scale this as a percentage.

4.3 EXPERIMENTAL SETUP

Each translation task was presented to the LLMs in a consistent way, containing the question stem, the translation examples, and the system prompt. The format of the question naturally lends itself to few-shot-prompting with the description of the task, then several lines (10 - 30) of example translations, followed by the actual question asking for a translation either from the constructed language to English or visa-via. Each translation was formatted in the same way throughout the dataset for consistency.

The system prompt was a simple but firm guide to further reinforce the ‘exam’ setting, and to encourage the model only to output the answer, not to re-frame the question in the response. This was designed not to enhance performance, we prefer to keep the core ‘prompting’ in the original question stem, and was mostly for controlling the output to encourage short answers of just the desired translation sentence.

This sort of output made it simpler to apply basic text cleaning to ensure we were evaluating the output translation and not any additional text. We decided this provided the cleanest analysis of the model’s ability to reason on the raw questions without additional guidance. Several system prompts were used, and the variation in model performance as a result was deemed negligible, so the system prompt was chosen based how clean the outputs were.

5 RESULTS

5.1 OVERALL PERFORMANCE COMPARISON

The core results across the suite of benchmarks are presented in Table 1, where the best result for each metric is shown in **bold** typeface, the second best in **orange**, and the third best in **blue**. Example results are shown in Fig. 3 for two metrics, ranked by the performance, full results in Appendix.A.2.

5.2 MODEL RANKING

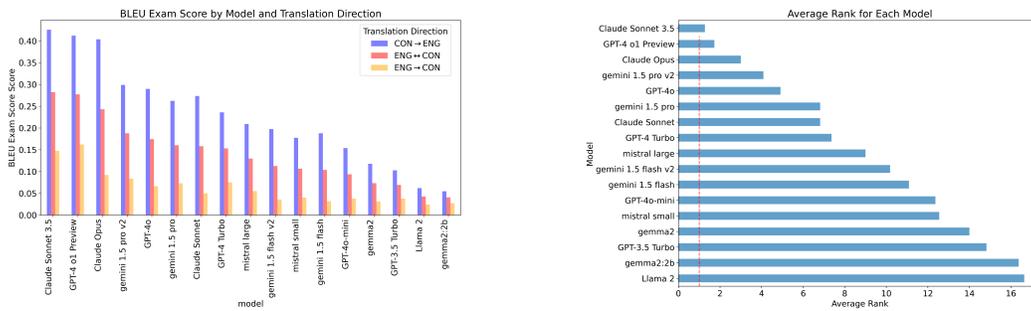
We can see clear stratification in terms of model performance, with Claude Sonnet 3.5 achieving the best results in 19/33 metrics, GPT4-o1 (preview) coming in second place with 14/33, and Claude Opus in third. The rankings are quantified in Table. 2, showing the average rank, and the number of Top-1, Top-3 and Top-5 placings for each model. This average rank performance is visualised in Fig.3b, where we can see clusters of models grouped together which have similar performance.

Table 1: Results shown for all models evaluated using all metrics. The scoring is broken down into the following categories: Evaluation method (BLEU, METEOR, etc.), translations in both directions, english to conlang, and conlang to english. These straight metrics are simple averages of the score for the translation weighting each question identically. The evaluation methods (where appropriate) have an exam setting reported, where the translation score is weighted by the points for the question, and the score over the entire question set reported. These scores are also broken into the three translation direction groups. This weighting gives more weight to more complex questions with longer prompts.

Score	Direction	Claude Opus	Claude Sonnet	Claude Sonnet 3.5	GPT-3.5 Turbo	GPT-4 Turbo	GPT-4 o1 Preview*	GPT-4o	GPT-4o Mini	Llama 2	Gemini 1.5 Flash	Gemini 1.5 Flash v2	Gemini 1.5 Pro	Gemini 1.5 Pro v2	Gemma 2	Gemma 2.2b	Mistral Large	Mistral Small	
BLEU	↑	ALL	0.249	0.162	0.296	0.073	0.169	0.288	0.187	0.100	0.045	0.112	0.116	0.169	0.196	0.078	0.042	0.139	0.111
BLEU	↑	CON→ENG	0.393	0.263	0.426	0.106	0.253	0.417	0.299	0.159	0.063	0.195	0.195	0.265	0.301	0.121	0.056	0.218	0.177
BLEU	↑	ENG→CON	0.097	0.055	0.158	0.038	0.079	0.164	0.067	0.039	0.026	0.033	0.035	0.076	0.086	0.032	0.028	0.056	0.040
BLEU Exam	↑	ALL	0.243	0.158	0.282	0.069	0.153	0.277	0.175	0.094	0.043	0.104	0.113	0.161	0.188	0.073	0.041	0.130	0.107
BLEU Exam	↑	CON→ENG	0.404	0.273	0.426	0.103	0.236	0.412	0.290	0.154	0.062	0.188	0.197	0.262	0.299	0.118	0.055	0.209	0.177
BLEU Exam	↑	ENG→CON	0.092	0.050	0.147	0.038	0.075	0.162	0.066	0.038	0.024	0.032	0.036	0.073	0.084	0.031	0.027	0.055	0.040
CharacTER	↓	ALL	0.366	0.427	0.320	0.555	0.440	0.332	0.394	0.511	2.637	0.495	0.476	0.441	0.402	0.554	1.507	0.467	0.548
CharacTER	↓	CON→ENG	0.280	0.357	0.235	0.493	0.386	0.303	0.346	0.431	1.124	0.423	0.401	0.350	0.318	0.505	2.230	0.413	0.500
CharacTER	↓	ENG→CON	0.457	0.501	0.409	0.620	0.498	0.360	0.445	0.593	4.239	0.563	0.553	0.529	0.489	0.606	0.730	0.524	0.599
chrF	↑	ALL	0.571	0.488	0.620	0.305	0.483	0.615	0.509	0.370	0.212	0.428	0.416	0.488	0.520	0.352	0.228	0.453	0.376
chrF	↑	CON→ENG	0.678	0.565	0.711	0.339	0.531	0.675	0.581	0.425	0.259	0.493	0.480	0.564	0.600	0.395	0.250	0.516	0.458
chrF	↑	ENG→CON	0.458	0.407	0.524	0.268	0.433	0.558	0.432	0.312	0.163	0.365	0.350	0.415	0.437	0.306	0.204	0.386	0.289
chrF Exam	↑	ALL	0.568	0.486	0.614	0.305	0.475	0.616	0.499	0.367	0.206	0.424	0.417	0.487	0.520	0.350	0.227	0.449	0.374
chrF Exam	↑	CON→ENG	0.686	0.572	0.713	0.338	0.524	0.675	0.578	0.418	0.257	0.493	0.484	0.567	0.605	0.394	0.248	0.510	0.462
chrF Exam	↑	ENG→CON	0.457	0.405	0.522	0.273	0.430	0.565	0.424	0.320	0.158	0.366	0.355	0.418	0.440	0.308	0.206	0.391	0.291
EED	↓	ALL	0.600	0.695	0.563	0.840	0.731	0.582	0.687	0.795	3.682	0.812	0.779	0.729	0.678	0.842	1.828	0.739	0.827
EED	↓	CON→ENG	0.434	0.541	0.398	0.698	0.579	0.461	0.539	0.636	1.223	0.626	0.589	0.555	0.494	0.705	2.474	0.591	0.706
EED	↓	ENG→CON	0.777	0.859	0.738	0.990	0.893	0.698	0.845	0.961	6.286	0.990	0.975	0.898	0.871	0.987	1.135	0.896	0.955
METEOR	↑	ALL	0.503	0.413	0.525	0.264	0.426	0.564	0.441	0.319	0.191	0.336	0.348	0.419	0.449	0.291	0.183	0.394	0.327
METEOR	↑	CON→ENG	0.725	0.622	0.740	0.396	0.612	0.726	0.650	0.485	0.277	0.546	0.550	0.636	0.661	0.443	0.256	0.600	0.514
METEOR	↑	ENG→CON	0.268	0.192	0.298	0.125	0.228	0.409	0.219	0.147	0.099	0.135	0.141	0.208	0.228	0.131	0.105	0.177	0.129
METEOR Exam	↑	ALL	0.487	0.397	0.506	0.250	0.403	0.550	0.421	0.301	0.178	0.321	0.340	0.402	0.436	0.279	0.172	0.378	0.316
METEOR Exam	↑	CON→ENG	0.732	0.629	0.741	0.386	0.602	0.724	0.646	0.473	0.268	0.541	0.556	0.636	0.665	0.437	0.248	0.597	0.518
METEOR Exam	↑	ENG→CON	0.258	0.180	0.285	0.123	0.216	0.402	0.209	0.142	0.094	0.133	0.144	0.202	0.220	0.131	0.099	0.171	0.126
ROUGE	↑	All	0.486	0.394	0.531	0.258	0.388	0.521	0.406	0.305	0.175	0.314	0.322	0.381	0.415	0.277	0.175	0.352	0.276
ROUGE	↑	CON→ENG	0.660	0.561	0.695	0.395	0.533	0.659	0.551	0.457	0.262	0.476	0.489	0.545	0.591	0.399	0.240	0.502	0.397
ROUGE	↑	ENG→CON	0.303	0.217	0.357	0.113	0.235	0.390	0.252	0.147	0.082	0.158	0.150	0.222	0.230	0.147	0.104	0.192	0.148
ROUGE Exam	↑	All	0.479	0.385	0.515	0.251	0.371	0.515	0.395	0.296	0.165	0.304	0.317	0.373	0.408	0.272	0.171	0.339	0.272
ROUGE Exam	↑	CON→ENG	0.671	0.569	0.696	0.393	0.524	0.659	0.547	0.452	0.255	0.475	0.494	0.547	0.595	0.403	0.239	0.498	0.400
ROUGE Exam	↑	ENG→CON	0.299	0.212	0.346	0.117	0.227	0.392	0.253	0.152	0.082	0.159	0.155	0.223	0.232	0.149	0.107	0.189	0.152
TER	↓	All	0.577	0.672	0.540	0.818	0.706	0.547	0.670	0.777	3.967	0.797	0.763	0.703	0.653	0.828	1.810	0.712	0.805
TER	↓	CON→ENG	0.397	0.499	0.360	0.669	0.538	0.412	0.508	0.608	1.223	0.597	0.559	0.518	0.454	0.680	2.440	0.547	0.668
TER	↓	ENG→CON	0.767	0.856	0.730	0.977	0.885	0.675	0.842	0.954	6.873	0.988	0.972	0.882	0.862	0.985	1.132	0.887	0.950

Table 2: The ranking results are presented in the following table. The average position of the model across all metrics (including accounting for those where the lowest result is best). Additionally the number of times the model ranks as the Top-1, in the Top-3 or in the Top-5 results is also given. All results are also broken into the translation directions.

score type	Direction	Claude Opus	Claude Sonnet	Claude Sonnet 3.5	GPT-3.5 Turbo	GPT-4 Turbo	GPT-4 o1 Preview*	GPT-4o	GPT-4o Mini	Llama 2	Gemini 1.5 Flash	Gemini 1.5 Flash v2	Gemini 1.5 Pro	Gemini 1.5 Pro v2	Gemma 2	Gemma 2.2b	Mistral Large	Mistral Small	
Average Rank	↓	All	3.000	6.818	1.273	14.818	7.364	1.727	4.909	12.364	16.636	11.091	10.182	6.818	4.091	14.000	16.364	9.000	12.545
Average Rank	↓	CON→ENG	2.273	6.091	1.000	14.545	8.000	2.727	5.364	12.545	16.000	10.727	10.091	6.545	4.000	14.000	17.000	9.182	12.909
Average Rank	↓	ENG→CON	3.091	7.545	2.000	14.182	5.636	1.000	5.182	11.545	17.000	12.091	11.364	7.182	4.727	16.000	8.636	12.091	
Top-1 Count	↑	All	0	0	8	0	0	3	0	0	0	0	0	0	0	0	0	0	0
Top-1 Count	↑	CON→ENG	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Top-1 Count	↑	ENG→CON	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0
Top-3 Count	↑	All	11	0	11	0	0	11	0	0	0	0	0	0	0	0	0	0	0
Top-3 Count	↑	CON→ENG	11	0	11	0	0	11	0	0	0	0	0	0	0	0	0	0	0
Top-3 Count	↑	ENG→CON	10	0	11	0	0	11	1	0	0	0	0	0	0	0	0	0	0
Top-5 Count	↑	All	11	0	11	0	0	11	11	0	0	0	0	0	11	0	0	0	0
Top-5 Count	↑	CON→ENG	11	3	11	0	0	11	8	0	0	0	0	0	11	0	0	0	0
Top-5 Count	↑	ENG→CON	11	2	11	0	7	11	5	0	0	0	0	8	0	0	0	0	0

432
433
434
435
436
437
438
439
440
441
442
443
444

(a) Ranked results of the BLEU Exam Score for each model. Shown for translation splits.

(b) Average ranking for each model shown in order, representing the top row of Table 2.

Figure 3: Comparison of BLEU Exam Scores and model rankings across all metrics.

5.3 TRANSLATION DIRECTION

One striking result we find is the discrepancy between metric scores and the translation direction. When translating from CON→ENG in all cases we see the model achieve a higher scores than in the ENG→CON cases. This reduction for ENG→CON can be seen clearly in the example metrics shown in Fig.3a. This is unsurprising, as we expect models trained on a largely English language corpus to be better at generating English text, but the scale of the discrepancy hints at a deeper issue. In Fig.3a showing the results for the BLEU exam score, the ENG→CON score of top performing model (Claude Sonnet 3.5) would rank 14th when compared to the CON→ENG . We argue this is the most challenging metric in the benchmark suite, and as seen in Table.1 is dominated by GPT-4 o1.

The tasks should be identical in both directions, however the model’s ability to generate grammatically correct English sentences perhaps offers an advantage, as a sentence like ‘*The judges hate the rams*’ while unusual follows an expected English pattern. If enough information can be aggregated from the original ‘*kertós misóntoi hógons*’ that it’s something about judges (probably plural), something about hate, something about rams (again probably plural). There aren’t very many ways to construct this as a grammatically correct English sentence. Additionally this is why the slightly eccentric nature of many of the translated phrases is such a desirable feature, if the model simply guesses at a sensible phrase but misses a subtle variation it will be penalised. This same phenomena is much harder to ape when doing translations of ENG→CON , and as such we find the ENG→CON BLEU exam score the strictest evaluation of the model’s reasoning ability.

5.4 PROMPT COMPLEXITY

Slightly counter to intuition, the model performance was not strongly inversely correlated with question complexity (as measured by points). We found some models dropped in performance with more complex questions, while others increased, and some seemed to find the medium level difficulty questions the hardest before rebounding with the most complex ones. Further details of these results can be found in Appendix. A.1, but we suggest the statistical uncertainty due to limited dataset size leaves this as an area for further study.

6 DISCUSSION

6.1 WINNERS AND LOSERS

As can be seen from Table.2 the top ranks are dominated by large and recent models - suggesting the convergence of large labs on a similar capability when it comes to reasoning. However, we see the Claude family of models represented in the 1st and 3rd ranked spots, with the new GPT-4 o1 (preview) model in 2nd. The multi-step reasoning has clearly had a big impact for this latest OpenAI model compared to previous iterations, and this is seen most apparently in the ENG→CON translations where this model placed 1st in 11/11 evaluations. Despite this posed as a step change in how LLMs will be deployed in the future, the two highest placed models from Anthropic are

486 strongly competitive and are apparently single models. This results raises further questions on what
487 differentiates the Anthropic dataset, or if there are similar chain-of-thought approaches being done
488 behind the scenes.

490 6.2 IMPLICATIONS FOR LLM CAPABILITIES

491
492 The questions in CLEAR aim to evaluate the ability of LLMs to parse new information and synthesis
493 new meanings and connections. As the exam is based on a constructed, and therefore unseen,
494 languages this task does not rely on memorisation, but on a basic grasp of language components,
495 and the ability to apply fairly simply logical patterns to those components using a broad set of
496 grammatical rules. The results are promising, especially taken for the CON→ENG direction, however
497 the disparity found in the ENG→CON translations suggest the current state of LLMs struggle to adapt
498 to new information and to update beliefs accordingly when the content falls outside of the training
499 corpus. This has implications for the more open ended reasoning tasks we would ultimately wish to
500 utilise LLMs for, and suggests significant further work is needed.

501 6.3 LIMITATIONS OF THE STUDY

502
503 A clear limitation is the dataset size, we have a dataset consisting of 140 unique questions, from 12
504 different exams, so there is a fair degree of overlap between questions. For example for the same
505 set of 10 example translations we will have 4 versions of the question stem, just with different final
506 translation.

507 Another major limitation is the evaluation of the impact tokenization plays on this particular task.
508 Conlangs by their nature often are composed of unusual strings of characters, and we suspect this
509 plays a role in some of the difficulty the LLMs demonstrate with the task. A full evaluation of the
510 impact of tokenization is beyond the scope of this paper.

511 A final major limitation of this study is linked to the power of this initial evaluation, we do not think
512 as of submission that these questions are included in any training set we know of. Publishing this
513 work and encouraging evaluation of the models on this dataset will no-doubt increase the availability
514 of these example questions / answers in web-crawled datasets, removing the power of this specific
515 study. Such dilution takes time however, and we hope this study will encourage further work in this
516 area.

517 6.4 FUTURE WORK

518
519 Future work will require the development of a significantly larger, and more diverse set of Conlangs,
520 however this requires a substantial effort from skilled linguists to ensure the Conlangs are both valid,
521 and the question structure consisting of the minimally required information to derive the translation.
522 There would also be value in exploring how this kind of in-context learning could be used to update
523 the properties of certain tokens, and how this might lead to simple methods to circumvent safety
524 features of current SOTA LLMs.

525 7 CONCLUSION

526
527 We have presented a new dataset benchmark, CLEAR, to evaluate the in-context reasoning capabil-
528 ities of LLMs in performing translation tasks on unseen Conlangs. Our initial benchmarks across a
529 diverse set of LLMs show promising abilities in the largest models to adapt to new information and
530 solve complex reasoning tasks. However, we observed a significant drop in performance for tasks
531 requiring a higher degree of reasoning, particularly in translating ENG→CON. This task, which
532 removes the possibility of relying on the model's ability to construct grammatically correct En-
533 glish sentences, exposes limitations in true reasoning capabilities. These findings suggest that while
534 progress has been made, there remains a substantial gap in general reasoning ability. We hope that
535 by presenting as a new public benchmark, we will encourage further research targeting this specific
536 problem. Ultimately, this work aims to pave the way towards models capable of robust, general
537 reasoning ability across diverse and novel contexts.

538
539

REFERENCES

- 540
541
542 GPT-4o system card | OpenAI. URL [https://openai.com/index/
543 gpt-4o-system-card/](https://openai.com/index/gpt-4o-system-card/).
- 544 Anthropic. Claude sonnet 3.5, 2024. URL [https://www.anthropic.com/news/
545 claude-3-5-sonnet](https://www.anthropic.com/news/claude-3-5-sonnet).
- 546
547 Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Im-
548 proved Correlation with Human Judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and
549 Clare Voss (eds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Mea-
550 sures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan, June
551 2005. Association for Computational Linguistics. URL [https://aclanthology.org/
552 W05-0909](https://aclanthology.org/W05-0909).
- 553 Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhari-
554 wal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal,
555 Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M.
556 Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin,
557 Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford,
558 Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. Technical Re-
559 port arXiv:2005.14165, arXiv, July 2020. URL <http://arxiv.org/abs/2005.14165>.
560 arXiv:2005.14165 [cs] type: article.
- 561 François Chollet. On the measure of intelligence. URL [http://arxiv.org/abs/1911.
562 01547](http://arxiv.org/abs/1911.01547).
- 563 Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam
564 Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh,
565 Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam
566 Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James
567 Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Lev-
568 skaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin
569 Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret
570 Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick,
571 Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica
572 Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Bren-
573 nan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas
574 Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. PaLM: Scaling Language Modeling with Pathways,
575 October 2022. URL <http://arxiv.org/abs/2204.02311>. arXiv:2204.02311 [cs].
- 576 Ollama Collaboration. Ollama, 2024. URL <https://github.com/ollama/ollama>.
- 577
578 Google Deepmind. Gemini pro, 2024. URL [https://deepmind.google/technologies/
579 gemini/pro/](https://deepmind.google/technologies/gemini/pro/).
- 580 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep
581 Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May 2019.
582 URL <http://arxiv.org/abs/1810.04805>. arXiv: 1810.04805.
- 583
584 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha
585 Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony
586 Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark,
587 Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere,
588 Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris
589 Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong,
590 Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny
591 Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino,
592 Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael
593 Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Ander-
son, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah
Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan

594 Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Ma-
595 hadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy
596 Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak,
597 Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Al-
598 wala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini,
599 Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yearly, Laurens van der
600 Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo,
601 Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Man-
602 nat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova,
603 Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal,
604 Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur
605 Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhar-
606 gava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong,
607 Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic,
608 Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sum-
609 baly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa,
610 Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang,
611 Sharath Rparathy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende,
612 Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney
613 Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom,
614 Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta,
615 Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petro-
616 vic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang,
617 Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuwei Wang, Yaelle Goldschlag, Yashesh Gaur,
618 Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre
619 Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha
620 Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand,
621 Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein,
622 Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples,
623 Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco,
624 Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman,
625 Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto
626 De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Bran-
627 don Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina
628 Mejia, Changan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai,
629 Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li,
630 Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana
631 Liskovich, Didem Foss, DingKang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil,
632 Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Ar-
633 caute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco
634 Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella
635 Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory
636 Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang,
637 Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Gold-
638 man, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman,
639 James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer
640 Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe
641 Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie
642 Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun
643 Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal
644 Chawla, Kushal Lakhota, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva,
645 Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian
646 Khabba, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson,
647 Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Ke-
neally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel
Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mo-
hammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navy-
ata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong,
Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli,

- 648 Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux,
649 Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao,
650 Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li,
651 Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott,
652 Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Sa-
653 tadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lind-
654 say, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang
655 Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen
656 Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho,
657 Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser,
658 Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Tim-
659 othy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan,
660 Vinay Satish Kumar, Vishal Mangla, Vitor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu
661 Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Con-
662 stable, Xiaocheng Tang, Xiaofang Wang, Xiaojuan Wu, Xiaolan Wang, Xide Xia, Xilun Wu,
663 Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi,
664 Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef
665 Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models. URL
666 <http://arxiv.org/abs/2407.21783>.
- 667 Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large
668 language models are zero-shot reasoners. URL <http://arxiv.org/abs/2205.11916>.
- 669 Gregor Leusch, Nicola Ueffing, and Hermann Ney. CDER: Efficient MT evaluation using block
670 movements. In Diana McCarthy and Shuly Wintner (eds.), *11th Conference of the European
671 Chapter of the Association for Computational Linguistics*, pp. 241–248. Association for Compu-
672 tational Linguistics. URL <https://aclanthology.org/E06-1031>.
- 673 Yuesen Li, Chengyi Gao, Xin Song, Xiangyu Wang, Yungang Xu, and Suxia Han. Drug-
674 GPT: A GPT-based strategy for designing potential ligands targeting specific proteins. URL
675 <https://www.biorxiv.org/content/10.1101/2023.06.29.543848v1>. Pages:
676 2023.06.29.543848 Section: New Results.
- 677 Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summa-
678 rization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational
679 Linguistics. URL <https://aclanthology.org/W04-1013>.
- 680 Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle
681 Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh
682 Koura, Vishrav Chaudhary, Brian O’Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva,
683 Mona Diab, Veselin Stoyanov, and Xian Li. Few-shot Learning with Multilingual Language Mod-
684 els, November 2022. URL <http://arxiv.org/abs/2112.10668>. arXiv:2112.10668
685 [cs].
- 686 Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. LogiQA: A
687 challenge dataset for machine reading comprehension with logical reasoning. URL <http://arxiv.org/abs/2007.08124>.
- 688
689
690 Chi-kiu Lo. YiSi - a Unified Semantic MT Quality Evaluation and Estimation Metric for Languages
691 with Different Levels of Available Resources. In Ondřej Bojar, Rajen Chatterjee, Christian Fe-
692 dermann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes,
693 Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névél, Mariana Neves,
694 Matt Post, Marco Turchi, and Karin Verspoor (eds.), *Proceedings of the Fourth Conference on
695 Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 507–513, Florence, Italy,
696 August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5358. URL
697 <https://aclanthology.org/W19-5358>.
- 698
699 Mistral. Mistral large model, 2024. URL <https://docs.mistral.ai>.
- 700 Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David
701 Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Au-
gustus Odena. Show your work: Scratchpads for intermediate computation with language models.
URL <http://arxiv.org/abs/2112.00114>.

- 702 OpenAI. Learning to reason with llms, 2024. URL [https://openai.com/index/](https://openai.com/index/learning-to-reason-with-llms/)
703 [learning-to-reason-with-llms/](https://openai.com/index/learning-to-reason-with-llms/).
704
- 705 OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Floren-
706 cia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red
707 Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Moham-
708 mad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher
709 Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brock-
710 man, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann,
711 Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis,
712 Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey
713 Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux,
714 Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila
715 Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix,
716 Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gib-
717 son, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan
718 Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hal-
719 lacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan
720 Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu,
721 Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun
722 Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Ka-
723 mali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook
724 Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel
725 Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen
726 Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel
727 Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez,
728 Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv
729 Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney,
730 Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick,
731 Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel
732 Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Ra-
733 jeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe,
734 Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel
735 Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe
736 de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny,
737 Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl,
738 Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra
739 Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders,
740 Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Sel-
741 sam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor,
742 Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky,
743 Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang,
744 Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Pre-
745 ston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vi-
746 jayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan
747 Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng,
748 Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Work-
749 man, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming
750 Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tian-
751 hao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 technical report. URL
752 <http://arxiv.org/abs/2303.08774>.
- 750 Joybrata Panja and Sudip Kumar Naskar. ITER: Improving Translation Edit Rate through Optimiz-
751 able Edit Costs. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette
752 Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz,
753 Matteo Negri, Aurélie Névélol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin
754 Verspoor (eds.), *Proceedings of the Third Conference on Machine Translation: Shared Task Pa-
755 pers*, pp. 746–750, Belgium, Brussels, October 2018. Association for Computational Linguistics.
doi: 10.18653/v1/W18-6455. URL <https://aclanthology.org/W18-6455>.

- 756 Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a Method for Automatic
757 Evaluation of Machine Translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.),
758 *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp.
759 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.
760 doi: 10.3115/1073083.1073135. URL <https://aclanthology.org/P02-1040>.
- 761 Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In Ondřej Bojar,
762 Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara
763 Logacheva, and Pavel Pecina (eds.), *Proceedings of the Tenth Workshop on Statistical Machine*
764 *Translation*, pp. 392–395, Lisbon, Portugal, September 2015. Association for Computational Lin-
765 guistics. doi: 10.18653/v1/W15-3049. URL <https://aclanthology.org/W15-3049>.
- 766 Maja Popović. chrF++: words helping character n-grams. In Ondřej Bojar, Christian Buck, Rajen
767 Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Ji-
768 meno Yepes, Philipp Koehn, and Julia Kreutzer (eds.), *Proceedings of the Second Conference*
769 *on Machine Translation*, pp. 612–618, Copenhagen, Denmark, September 2017. Association for
770 Computational Linguistics. doi: 10.18653/v1/W17-4770. URL <https://aclanthology.org/W17-4770>.
- 771 Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi
772 Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified
773 Text-to-Text Transformer, September 2023. URL <http://arxiv.org/abs/1910.10683>.
774 arXiv:1910.10683 [cs, stat].
- 775 Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A Neural Framework for MT
776 Evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the*
777 *2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2685–
778 2702, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/
779 2020.emnlp-main.213. URL <https://aclanthology.org/2020.emnlp-main.213>.
- 780 David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien
781 Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a
782 benchmark. URL <http://arxiv.org/abs/2311.12022>.
- 783 Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen, Yossi
784 Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton,
785 Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez,
786 Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and
787 Gabriel Synnaeve. Code llama: Open foundation models for code.
- 788 Lingfeng Shen, Aayush Mishra, and Daniel Khashabi. Do pretrained Transformers Learn In-
789 Context by Gradient Descent?, June 2024. URL <http://arxiv.org/abs/2310.08540>.
790 arXiv:2310.08540 [cs].
- 791 Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. A study of
792 translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the*
793 *Association for Machine Translation in the Americas: Technical Papers*, pp. 223–231. Associa-
794 tion for Machine Translation in the Americas. URL <https://aclanthology.org/2006.amta-papers.25>.
- 795 Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam
796 Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska,
797 Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Ko-
798 curek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda
799 Askeel, Amanda Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea San-
800 tilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang,
801 Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh,
802 Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov,
803 Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan
804 Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam He-
805 dayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin,
806 Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri

810 Ramrez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris
811 Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy
812 Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea,
813 Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel
814 Khashabi, Daniel Levy, Daniel Mosegu Gonzlez, Danny Hernandez, Danqi Chen, Daphne
815 Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep
816 Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hup-
817 kes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekate-
818 rina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth
819 Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem,
820 Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Ev-
821 genii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martnez-Plumed, Francesca Happ,
822 Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germn
823 Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-
824 Lpez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh
825 Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schtze, Hiromu Yakura, Hong-
826 ming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson
827 Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernandez Fisac, James B. Simon, James Koppel,
828 James Zheng, James Zou, Jan Kocon, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Sohl-
829 Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer
830 Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian
831 Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jorg Frohberg,
832 Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum,
833 Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrish-
834 nan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory
835 Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richard-
836 son, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando,
837 Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng
838 He, Luis Oliveros Coln, Luke Metz, Lutfi Kerem enel, Maarten Bosma, Maarten Sap, Maartje
839 ter Hoeve, Madotto Andrea, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan,
840 Marco Marelli, Marco Maru, Maria Jose Ramrez Quintana, Marie Tolkieln, Mario Giulianelli,
841 Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Matys Schubert, Med-
842 ina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen,
843 Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Micha Swedrowski, Michele
844 Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimeo Xu, Mirac Suzgun, Mo Tiwari,
845 Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdah Gheini, Mukund Varma T, Nanyun Peng,
846 Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick
847 Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S.
848 Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi,
849 Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu
850 Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eck-
851 ersley, Phu Mon Htut, Pinyu Hwang, Piotr Mikowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli,
852 Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel,
853 Rahel Habacker, Ramn Risco Delgado, Raphael Millire, Rhythm Garg, Richard Barnes, Rif A.
854 Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Ro-
855 man Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov,
856 Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Moham-
857 mad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R.
858 Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghaz-
859 arian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schus-
860 ter, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar
861 Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upad-
862 hyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy,
863 Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislav Dehaene,
Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Pianta-
dosi, Stuart M. Shieber, Summer Mishnerghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen,
Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Theo Desbordes, Theodore
Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Timo-
thy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar

- 864 Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak,
865 Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus,
866 William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinyi Wu, Xudong
867 Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi
868 Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary
869 Seid, Zhao Xinran, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond
870 the imitation game: Quantifying and extrapolating the capabilities of language models. URL
871 <http://arxiv.org/abs/2206.04615>. type: article.
- 872 Peter Stanchev, Weiyue Wang, and Hermann Ney. EED: Extended Edit Distance Measure for Ma-
873 chine Translation. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette
874 Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins,
875 Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and
876 Karin Verspoor (eds.), *Proceedings of the Fourth Conference on Machine Translation (Volume 2:
877 Shared Task Papers, Day 1)*, pp. 514–520, Florence, Italy, August 2019. Association for Com-
878 putational Linguistics. doi: 10.18653/v1/W19-5359. URL [https://aclanthology.org/
879 W19-5359](https://aclanthology.org/W19-5359).
- 880
881 Oyvind Tafjord, Bhavana Dalvi Mishra, and Peter Clark. ProofWriter: Generating implications,
882 proofs, and abductive statements over natural language. URL [https://arxiv.org/abs/
883 2012.13048v2](https://arxiv.org/abs/2012.13048v2).
- 884
885 Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya
886 Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard
887 Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex
888 Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, An-
889 tonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo,
890 Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric
891 Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Hen-
892 ryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski,
893 Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu,
894 Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee,
895 Lucas Dixon, Machel Reid, Maciej Mikula, Mateo Wirth, Michael Sharman, Nikolai Chinaev,
896 Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko
897 Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo
898 Liu, Ryan Mullins, Samuel L. Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree
899 Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech
900 Stokowiec, Yu-hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh
901 Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin
902 Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah
903 Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on
904 gemini research and technology. URL <http://arxiv.org/abs/2403.08295>.
- 905
906 Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée
907 Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Ar-
908 mand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation
909 language models. URL <http://arxiv.org/abs/2302.13971>.
- 910
911 Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don’t always
912 say what they think: Unfaithful explanations in chain-of-thought prompting. URL [http://
913 arxiv.org/abs/2305.04388](http://arxiv.org/abs/2305.04388).
- 914
915 Karthik Valmeekam, Kaya Stechly, and Subbarao Kambhampati. LLMs still can’t plan; can LRMs?
916 a preliminary evaluation of OpenAI’s o1 on PlanBench. URL [http://arxiv.org/abs/
917 2409.13373](http://arxiv.org/abs/2409.13373).
- 918
919 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez,
920 Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. URL [http://arxiv.org/
921 abs/1706.03762](http://arxiv.org/abs/1706.03762).

- 918 Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. CharacTer: Translation
919 Edit Rate on Character Level. In Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Fed-
920 ermann, Liane Guillou, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Aurélie Névél, Mariana Neves, Pavel Pecina, Martin Popel, Philipp Koehn, Christof Monz, Matteo Negri, Matt
921 Post, Lucia Specia, Karin Verspoor, Jörg Tiedemann, and Marco Turchi (eds.), *Proceedings of the*
922 *First Conference on Machine Translation: Volume 2, Shared Task Papers*, pp. 505–510, Berlin,
923 Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2342.
924 URL <https://aclanthology.org/W16-2342>.
925
- 926 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc
927 Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models.
928 URL <https://arxiv.org/abs/2201.11903v6>.
929
- 930 BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić,
931 Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé,
932 Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Amman-
933 manchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji
934 Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lu-
935 cile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite,
936 Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Al-
937 ham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou,
938 Chris Emezue, Christopher Klam, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani,
939 Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan,
940 Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza
941 Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier
942 de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing,
943 Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon
944 Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz,
945 Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh,
946 Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subra-
947 mani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo
948 Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rhea Harliman,
949 Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian
950 Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, So-
951 maieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo
952 Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lep-
953 ercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei
954 Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesh Sharma,
955 Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan
956 Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao,
957 Lintang Sutawika, M. Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Tee-
958 han, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers,
959 Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong,
960 Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung,
961 Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared
962 Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myr-
963 iam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre
964 Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden
965 Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anas-
966 tasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névél, Charles Lovering,
967 Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli
968 Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova,
969 Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat,
970 Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal,
971 Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shav-
rina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav
Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice
Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony
Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh Haji-
Hosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel

- 972 McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Ed-
973 ward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezane-
974 jad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse
975 Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Mar-
976 got Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed
977 Ghauri, Mykola Burynek, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanre-
978 waju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas
979 Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach
980 Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, An-
981 tonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou,
982 Chirag Jain, Chuxin Xu, Clémentine Fourier, Daniel León Perrián, Daniel Molano, Dian Yu,
983 Enrique Manjavacas, Fabio Barth, Florian Fuhrmann, Gabriel Altay, Giyaseddin Bayrak, Gully
984 Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde,
985 Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shin-
986 zato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A. Castillo, Mari-
987 anna Nezhurina, Mario Sängler, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel
988 De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan
989 Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya
990 Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel
991 Cahyawijaya, Samuele Garda, Shlok S. Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott,
992 Sinee Sang-aaronsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gi-
993 gant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman,
994 Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada,
995 and Thomas Wolf. BLOOM: A 176b-parameter open-access multilingual language model. URL
996 <http://arxiv.org/abs/2211.05100>.
- 996 Xi Ye and Greg Durrett. The unreliability of explanations in few-shot prompting for textual reason-
997 ing. URL <http://arxiv.org/abs/2205.03401>.
- 998 Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping Reasoning With
999 Reasoning, May 2022. URL <http://arxiv.org/abs/2203.14465>. arXiv:2203.14465
1000 [cs].
- 1001 Eric Zelikman, Georges Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah D. Goodman.
1002 Quiet-STaR: Language Models Can Teach Themselves to Think Before Speaking, March 2024.
1003 URL <http://arxiv.org/abs/2403.09629>. arXiv:2403.09629 [cs].
- 1004
- 1005 Di Zhang, Xiaoshui Huang, Dongzhan Zhou, Yuqiang Li, and Wanli Ouyang. Accessing GPT-4
1006 level mathematical olympiad solutions via monte carlo tree self-refine with LLaMa-3 8b. URL
1007 <http://arxiv.org/abs/2406.07394>.
- 1008
- 1009 Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Eval-
1010 uating Text Generation with BERT, February 2020. URL <http://arxiv.org/abs/1904.09675>.
1011 arXiv:1904.09675 [cs].
- 1012
- 1013
- 1014
- 1015
- 1016
- 1017
- 1018
- 1019
- 1020
- 1021
- 1022
- 1023
- 1024
- 1025

1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079

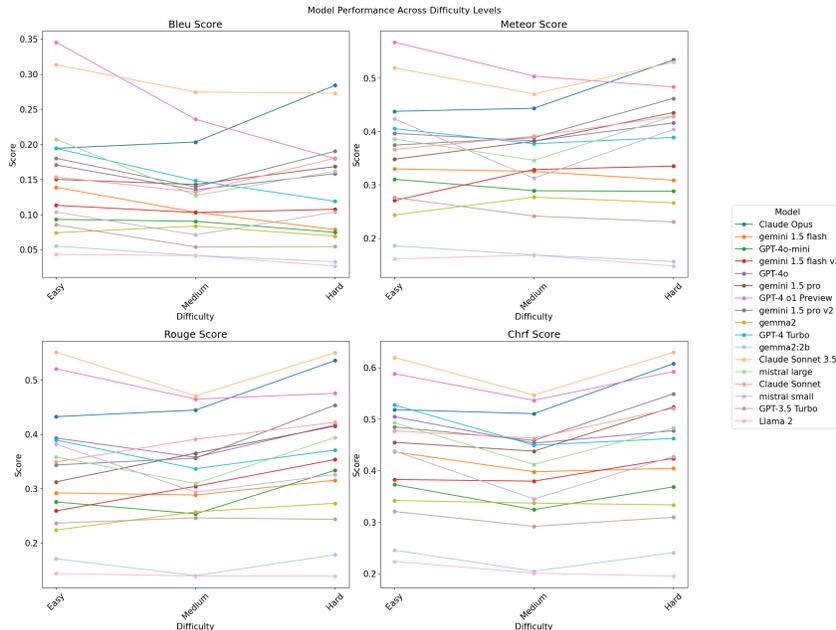


Figure 4: Plotting the relative performance for each model stratified by question difficulty - as measured by the number of points associated with the questions.

A APPENDIX

A.1 BY QUESTION DIFFICULTY

We investigated the performance of the models against the question difficulty, as measured by the proxy of number of points per question. The results as shown in Fig. 4 and in Table.3 there are some differences, but the results are not statistically significant considering the relatively small number of questions in each category once broken into the three difficulties. We can see however that there is not just a simple decrease in performance with complexity, the complexity is at least counteracted by the additional examples in the prompt.

Model	BLEU Score					METEOR Score					ROUGE Score					CHRF Score				
	Easy	Medium	Hard	%Δ	%Δ	Easy	Medium	Hard	%Δ	%Δ	Easy	Medium	Hard	%Δ	%Δ	Easy	Medium	Hard	%Δ	%Δ
Claude Opus	0.194	0.203	0.284	4.6%	46.4%	0.438	0.443	0.534	1.1%	22.0%	0.433	0.445	0.536	2.8%	23.8%	0.518	0.510	0.607	-1.5%	17.2%
Claude Sonnet	0.153	0.132	0.180	-13.7%	17.6%	0.366	0.391	0.429	6.8%	17.2%	0.349	0.391	0.422	12.0%	20.9%	0.477	0.463	0.521	-2.9%	9.2%
Claude Sonnet 3.5	0.313	0.275	0.273	-12.2%	-12.8%	0.519	0.470	0.530	-9.4%	2.1%	0.551	0.471	0.550	-14.5%	-0.2%	0.619	0.547	0.629	-11.6%	1.6%
GPT-4 Turbo	0.085	0.054	0.054	-36.6%	-36.5%	0.276	0.241	0.231	-12.6%	-16.3%	0.236	0.246	0.244	4.1%	3.3%	0.321	0.292	0.309	-9.0%	-3.7%
GPT-4o	0.194	0.148	0.119	-23.7%	-38.8%	0.405	0.377	0.389	-6.9%	-4.0%	0.390	0.337	0.371	-13.6%	-4.9%	0.527	0.449	0.462	-14.8%	-12.3%
GPT-4o1 Preview	0.345	0.236	0.180	-31.7%	-47.9%	0.567	0.504	0.483	-11.1%	-14.8%	0.521	0.465	0.475	-10.7%	-8.8%	0.588	0.536	0.592	-8.8%	0.7%
GPT-4o mini	0.171	0.135	0.158	-21.0%	-7.5%	0.396	0.382	0.416	-3.5%	5.0%	0.393	0.358	0.417	-8.9%	6.1%	0.504	0.453	0.477	-10.1%	-5.4%
Llama 2	0.093	0.090	0.075	-3.2%	-19.4%	0.310	0.289	0.288	-6.8%	-7.1%	0.276	0.254	0.334	-8.0%	21.1%	0.373	0.324	0.368	-13.1%	-1.3%
Gemma2	0.043	0.041	0.027	-4.2%	-37.4%	0.161	0.168	0.148	4.4%	-8.1%	0.144	0.139	0.139	-3.3%	-3.3%	0.224	0.201	0.195	-10.3%	-12.8%
Gemini 1.5 flash	0.139	0.103	0.078	-25.6%	-43.5%	0.330	0.325	0.309	-1.5%	-6.4%	0.292	0.288	0.316	-1.3%	8.2%	0.436	0.398	0.404	-8.7%	-7.3%
Gemini 1.5 flash v2	0.113	0.103	0.107	-8.9%	-5.3%	0.271	0.329	0.335	21.5%	23.7%	0.259	0.304	0.354	17.4%	36.6%	0.383	0.379	0.424	-1.0%	10.7%
Gemini 1.5 pro	0.150	0.143	0.168	-4.7%	12.2%	0.348	0.381	0.435	9.5%	25.1%	0.313	0.365	0.415	16.7%	32.6%	0.455	0.438	0.524	-3.7%	15.2%
Gemini 1.5 pro v2	0.180	0.139	0.190	-22.7%	5.5%	0.374	0.388	0.462	3.7%	23.5%	0.344	0.356	0.454	3.5%	32.0%	0.485	0.458	0.549	-5.6%	13.2%
Gemma2	0.074	0.084	0.069	13.5%	-6.6%	0.243	0.277	0.266	13.9%	9.4%	0.224	0.257	0.273	14.7%	21.9%	0.342	0.337	0.333	-1.5%	-2.6%
Gemma2.2b	0.055	0.042	0.033	-23.7%	-40.4%	0.186	0.169	0.156	-9.2%	-15.9%	0.171	0.140	0.178	-18.1%	-4.3%	0.246	0.205	0.241	-16.7%	-2.0%
Mistral large	0.207	0.127	0.162	-38.7%	-21.7%	0.386	0.346	0.429	-10.3%	11.1%	0.358	0.310	0.394	-13.4%	10.0%	0.493	0.412	0.483	-16.4%	-2.0%
Mistral small	0.104	0.071	0.103	-31.7%	-0.2%	0.423	0.312	0.404	-26.2%	-4.5%	0.382	0.293	0.326	-23.3%	-14.7%	0.438	0.345	0.427	-21.2%	-2.5%

Table 3: Comparison of model performance across difficulty levels Easy, Medium, Hard for various metrics. Percentages in parentheses show the relative change from Easy to Medium and Easy to Hard levels, respectively.

A.2 ALL METRICS

We show here in Fig.5 the full set of metrics for all models split into translation directions. The same pattern can be observed across all scores with the top three models in very close competition, and with ENG→CON translations much more successfully conducted than ENG→CON .

1080
 1081
 1082
 1083
 1084
 1085
 1086
 1087
 1088
 1089
 1090
 1091
 1092
 1093
 1094
 1095
 1096
 1097
 1098
 1099
 1100
 1101
 1102
 1103
 1104
 1105
 1106
 1107
 1108
 1109
 1110
 1111
 1112
 1113
 1114
 1115
 1116
 1117
 1118
 1119
 1120
 1121
 1122
 1123
 1124
 1125
 1126
 1127
 1128
 1129
 1130
 1131
 1132
 1133

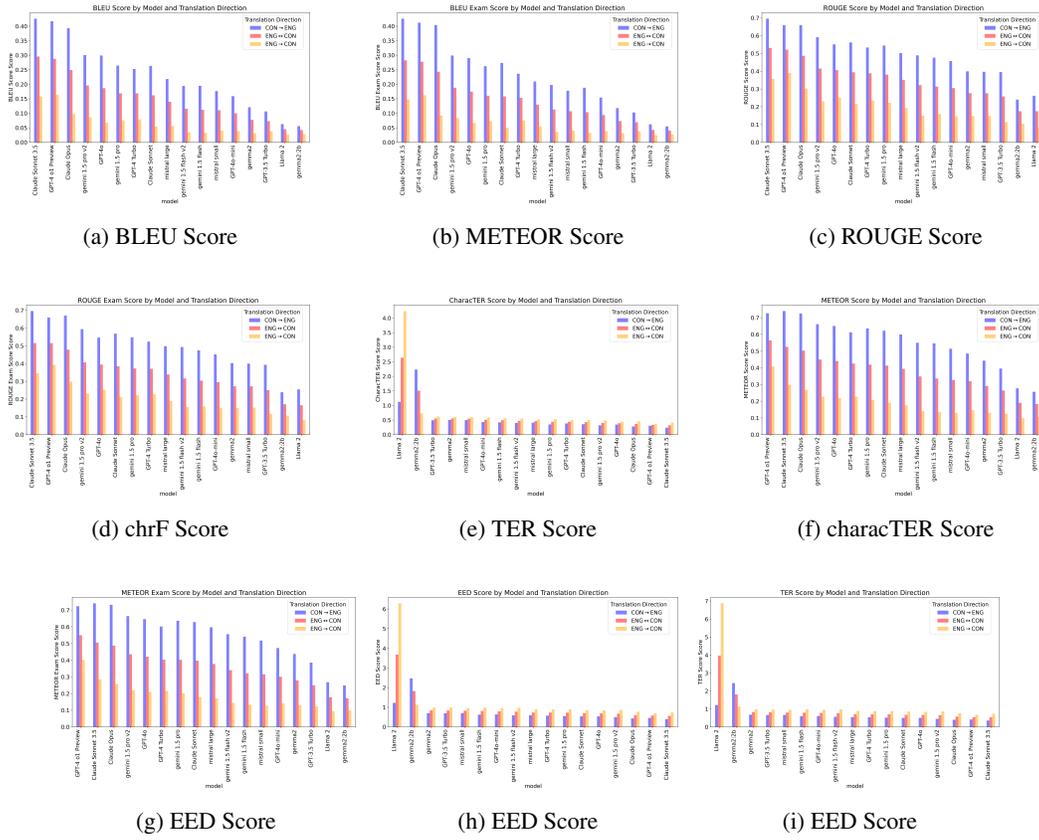


Figure 5: Comparison of Various Metric Scores by Model and Translation Direction