Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3056–3068
Marseille, 20-25 June 2022

On the Multilingual Capabilities of Very Large-Scale English Language Models

Jordi Armengol-Estapé*, Ona de Gibert Bonet*, Maite Melero
Barcelona Supercomputing Center
Plaça Eusebi Güell 1-3, Barcelona 08034, Spain
{jordi.armengol, ona.degibert, maite.melero}@bsc.es

Abstract

Generative Pre-trained Transformers (GPTs) have recently been scaled to unprecedented sizes in the history of machine learning. These language models have been shown to exhibit outstanding zero, one, and few-shot learning capabilities in a number of different tasks. Nevertheless, aside from anecdotal experiences, little is known regarding their multilingual capabilities, given the fact that the pre-training corpus is almost entirely composed of English text. In this work, we investigate its potential and limits in three tasks: extractive Question-Answering, text summarization and natural language generation for five different languages, as well as the effect of scale in terms of model size. Our results show that GPT-3 can be used, not only as a powerful generative pre-trained model for English, but for other languages as well, even for some with very few data in the training corpora, with room for improvement if optimization of the tokenization is addressed.

Keywords: Multilingual, Cross-lingual, Language Modeling

1. Introduction

Improving Natural Language Understanding (NLU) and Generation (NLG) by pre-training autoregressive language models based on the Transformer (Vaswani et al., 2017) decoder architecture has been commonplace since the original GPT (Generative Pretrained Transformer) (Radford and Narasimhan, 2018) first appeared. In the race to scale up these language models (Radford et al., 2019), the arrival of GPT-3 (Brown et al., 2020) has changed the rules of the game. As claimed by their creators, its ability to learn from a few examples "via text interaction" makes it stand out from the rest. Its impressive generative capabilities have caused a big sensation, not only at research level but also in the mainstream media.

A particular feature of GPT-3 is, besides the sheer size of the data it has been trained on, the fact that, although the data is generally of good quality, it has not been filtered for language (in purpose). Therefore, although GPT-3 is in principle a language model for English, its training data contains many other languages,[1] even if they account for a small portion of the dataset in comparison to English (93% by word count). Intuitively, one would expect that this quantity would not be enough to obtain a high-quality language model in these other languages, especially in the low-resource ones. Some evidence in this regard is provided by the large amount of data required to train language-specific models (Nozza et al., 2020). Even the multilingual ones[2] such as mBERT (Devlin et al., 2018) or XLM-R (Conneau et al., 2019) employ large multilingual datasets based on Wikipedia or CommonCrawl.[3]

In this work, we investigate the multilingual skills of different sizes of the well-known GPT-3 model, with a focus on how well this model scales in the different tasks for the different languages. While we devise a thorough evaluation, the main goal of this work is not to conduct *the ultimate* evaluation, due to current limitations that we will describe later on, but to open the research area at the intersection of language modeling, cross-lingual transfer and model scaling taking into account our findings. In summary, our contributions are as follows:

- We propose the new research area of studying the scale of the multilingual capabilities of (mostly) monolingual (English) language models.
- We conduct a thorough evaluation in both NLU and NLG for GPT-3, both with human and automatic evaluation, with a focus on the importance of model size, and describing the current limitations for evaluating the multilingual capabilities of GPT models
- We release to the community the outputs obtained from GPT-3 as a new resource for evaluating the multilingual skills of GPT-3-like models.[4] Apart from the model outputs, we will publicly release the code for the sake of reproducibility in the same repository.

The remaining of this article is organized as follows. In Section 2, we summarize the relevant related work. In

*Equal contribution.

[1] https://github.com/openai/gpt-3/tree/master/dataset_statistics

[2] Note that both mBERT and XLM-R are encoder-based models, unlike GPT, but the point still holds.

[3] https://commoncrawl.org/

[4] https://github.com/TeMU-BSC/gpt3-queries

Section 3, we propose our methodology and describe the experiments. Then, in Section 4, we go through the results of the experiments. Finally, in sections 5 and 6, we discuss the obtained results and propose future work.

2. Related Work

In Brown et al. (2020), the authors of GPT-3 already conducted a thorough evaluation in many different benchmarks, including Question-Answering, close tasks, and Natural Language Inference (NLI), among many others. They train and evaluate models of different sizes, and find that by simply scaling up the exact same architecture, the diminishing returns that one would expect are not observed. Recently, some works have estimated the increase in performance of autoregressive models in terms of model size, data, and compute (Kaplan et al., 2020), (Henighan et al., 2020). Also in Brown et al. (2020), and relevant to our work, authors evaluate GPT-3 in machine translation (MT), both in zero and few-shot settings, and find that in the latter, GPT-3 outperforms previous unsupervised neural MT models by 5 BLEU in some pairs. Specifically, this success is observed in the evaluated pairs in which English is the target language and not in the ones in which English is the source one, being GPT-3 an English language model. No other analysis involving languages other than English was conducted.

Since the original article of GPT-3, several works have investigated the capabilities and limits of the model in English (Zhao et al., 2021). Moreover, with the possibility of querying the model via API, hundreds of researchers, journalists and curious alike have embarked on all sorts of experiments, including automatic programming or solving arithmetic operations (Floridi and Chiriatti, 2020). The Internet is full of examples of the amazing generative capabilities of the model, from poetry, news or essay writing (Elkins and Chun, 2020).

Furthermore, many researchers are interested in the ethical concerns regarding such a capable generative model and are studying the impact it may had if it was released to the public (Dale, 2021; McGuffie and Newhouse, 2020). In a more consequential approach, with the purpose of harnessing the full learning potential of GPT, we are seeing the emergence of a new line of research exploring optimal ways to "prompt" the model (Liu et al., 2021).

Nevertheless, to our knowledge, no work has studied its potential for solving tasks in languages other than English, aside from machine translation. In this work, we investigate the multilingual skills of different size variants of the GPT-3 model.

3. Methodology

In this work we have explored how good GPT-3 is at understanding and generating natural language in different languages. We evaluate GPT-3's zero-shot multilingual capabilities on three generative tasks: Question Answering, Text Summarization, and "pure"[5] Text Generation.

In order to test the multilingual capabilities of the English-trained model, we have chosen five different languages. Four of them meet the double requirement of being present in the two multilingual datasets of reference to evaluate Question Answering and Summarization, namely XQuAD (Artetxe et al., 2019) and MLSUM (Scialom et al., 2020). Those languages are: German, Spanish, Russian, and Turkish. The fifth language that we have chosen is Catalan, which is a moderately under-resourced language, with very little data in the GPT-3 training corpus. To evaluate Catalan we have used a recently created Catalan translation of XQuAD (Armengol-Estapé et al., 2021) and a built-for-purpose summarization corpus. We have also included English in the evaluation for reference.[6]

To study the effect of scale, we run each experiment using the 4 engines provided in OpenAI's API,[7] in increasing size[8] (in parameters): Ada, Babbage, Curie, and Davinci.

Our hypothesis is that GPT-3, which is a mostly monolingual English model, will perform at levels close to those it obtains for English when evaluated on different languages.

3.1. Zero-shot Multilingual Question-Answering

Question-Answering (Q&A) refers to the task where, given a context and a question, the model must produce an answer. This task is usually approached as an extractive task with the answer located in the text by performing a per-token binary classification (i.e. decide whether the token is supposed to be part of the answer or not) on top of the model embeddings. However, we approach it by prompting a generative model, which makes the task more challenging. To evaluate the ability of GPT-3 at answering questions in several languages, we use XQuAD (Artetxe et al., 2019) a benchmark dataset for evaluating cross-lingual question answering performance. The dataset consists of a subset of 240 paragraphs and 1190 question-answer pairs from the development set of SQuAD v1.1 (Rajpurkar et al., 2016) together with their professional translations into ten languages. For Catalan, we use a recently published Catalan translation (Armengol-Estapé et al., 2021) of the same corpus. The fact that the corpus is

[5] Here we avoid the term "unconditional" because we need to condition the model to write in the desired language.

[6] The respective percentage of these languages in the GPT-3 training corpus in number of words is: English 92,647%; German 1,469%; Spanish 0,772%; Russian 0,188%; Turkish 0,059%; Catalan 0,017%

[7] https://beta.openai.com/

[8] To the best of our knowledge, OpenAI has not clarified the exact size of each of the models in the API. However, we use this estimation: https://blog.eleuther.ai/gpt3-model-sizes/

	CA	DE	EN	ES	RU	TR
Summarization	2.13	2.43	1.23	1.98	-	3.61
Question Answering	2.12	2.68	1.29	2.06	7.96	3.66

Table 1: Average tokens per word per language

entirely parallel across all languages excludes bias due to different degrees of complexity in the datasets.
For the evaluation, GPT-3 is prompted to answer one question at a time, pieced with its context as shown below (in bold, GPT-3's answer):

> *This is a Question-Answering system in English.*
>
> *Context: The Panthers defense gave up just 308 points [...]*
>
> *Question: How many points did the Panthers defense surrender?*
>
> *Answer:* ***308***

The whole prompt, including the instruction to answer the question (the first sentence), the context, the question, and the final word (*Answer:*) are written in the language that is being evaluated, with the hope that this will further condition the model to answer in the corresponding language (e.g., when evaluating the Spanish XQuAD, the whole prompt is written in Spanish). As sampling parameters, we use a temperature of 0 (since it is an extractive Question-Answering task, we do not want the model to be "creative"), a frequency penalty of 0 and a presence penalty of 0, with `top_p` = 0.95. We set `max_tokens` to 128 to allow the longest answers, which in XQuAD are generally short.
We consider the answer of the model to be the text immediately following the last word of the prompt ("Answer:"), cutting it when a new line is encountered (which generally means the model is starting to generate a new question). To evaluate the models, we have applied the evaluation script from SQuAD,[9] where articles are removed (in the case of English, German, Spanish, and Catalan). We compute both unlemmatized and lemmatized F1[10] and exact match (EM) scores.

3.2. Zero-shot Multilingual Text Summarization

Text Summarization is the task of producing a shorter version of a text while preserving the most relevant pieces of information. To evaluate it, we use MLSUM (Scialom et al., 2020), a large-scale multilingual summarization dataset obtained from online newspapers. In contrast to XQuAD, the multilingual content is not parallel, which may add noise to the comparison between languages. For English, we use the CNN/Daily Mail dataset (Hermann et al., 2015), which consists of online CNN and Daily Mail news articles. Since Catalan is not present in the original MLSUM, we have used another summarization dataset for Catalan, CaSum.[11] This dataset was built using a similar methodology, i.e. pairing news articles from the Agencia Catalana de Noticies[12] (an online news outlet) with their headlines and titles. More details about this dataset are provided in the Appendix.
Due to resource constraints, we randomly sampled a subset of 500 articles+summary from each summarization dataset. Before sampling, we applied two filters based on length and quality of the instances. Since GPT-3 has a context window of 2048 and we needed some margin to include the instruction and to allow the model to diverge from the ground truth, we discarded instances in which the concatenation of the article plus the summary was longer than 2000 tokens (using GPT-3's tokenizer). Then, since the quality of MLSUM is uneven, we used existing summarization models[13] to filter out summaries with a ROUGE score(Lin, 2004) below 0.1. Russian had to be entirely discarded for this experiment because the English-centric tokenization of GPT-3 articles produced tokenizations of single articles that did not fit in the context window of 2048 tokens. In the case of Catalan, since the dataset was manually validated, we did not apply any further quality filter.
As for Q&A, we also tested a zero-shot setting[14] where GPT-3 is asked to summarize one text at a time. Similarly to Nikolich and Puchkova (2021), we have used each text as prompt, together with the statement *TL;DR*, which stands for *Too long; didn't read* and is usually employed in online forums to indicate a summarization of the preceding text. Initially, we tried a prompt similar to the one we used in the case of Question-Answering,[15] with poor results.
Regarding the sampling parameters, as for Question-Answering, we used a temperature of 0.0 to prevent the model from being "creative", and a presence penalty of 0. This time, though, we expanded the maximum number of tokens to the maximum context window of

[9] `https://github.com/allenai/bi-att-flow/blob/master/squad/evaluate-v1.1.py`

[10] F1 score is the harmonic mean of precision and recall

[11] Citation metadata pending at the time of submitting the article. It will be added in the final version of the article.

[12] `https://www.acn.cat/`

[13] See "Models trained or fine-tuned on MLSUM" in `https://huggingface.co/datasets/mlsum`

[14] Note that in this case both few or one-shot learning settings would be generally impossible due to the limited context window of GPT-3.

[15] "This is a summarization system [...] Article: [...] Summary:"

GPT-3, 2048, and used a frequency penalty of 2, to encourage the models not to literally repeat the original text (which happened in some cases in the preliminary experiments we did).

We considered the answer of the model to be the text immediately following the last word of the prompt ("TL;DR:"), using a sentence splitter and taking the first 3 sentences of the output.[16] In case of a line break within these first 3 sentences, we considered only the previous sentences to the line break as the answer. Then, we normalized the punctuation.

Like any language generation tasks, abstractive summarization is challenging to evaluate (Sai et al., 2020). Aside from the difficulty of correctly evaluating a summary perhaps conveying the same meaning but using different words than the ones in the ground truth summary, the task itself is loosely defined, because it is even unclear to humans what length a summary should be. In supervised learning, the target summaries can be expected to be similar to those in the training set, but in zero-shot settings such as here, the model has no clue regarding the kind of summarization it must do.

With this in mind, we have included alternative automatic metrics apart from the standard one, ROUGE (based on N-gram co-occurrences), which has well-known caveats (Schluter, 2017). Specifically, we provide METEOR (Banerjee and Lavie, 2005) as well. In addition, we have included a human evaluation for a subset of summaries in two of the languages, English and Catalan, in which a pool of 3 evaluators were asked to rank summaries resulting from the 4 models plus the ground truth.

3.3. Multilingual Text Generation

Another way to evaluate the quality of the generative capabilities of GPT-3 is to submit the generated output to the Turing test, i.e. ask native evaluators if, according to them, a given sentence has been produced by a human or by artificial intelligence. Due to the high costs of this human evaluation, we have limited the test only to Catalan, and English for reference.

To obtain the synthetic sentences, we randomly sampled 20 articles both from the CNN/Daily Mail dataset and the CaSum dataset and used the obtained headlines as prompts to encourage the models to generate new text in the same language as the headlines.[17] As sampling parameters, this time we used a temperature of 0.7 (with `top_k=0.95`) to let the model be more "creative". We set `max_tokens` to 512, and both frequency penalty and presence penalty to 0. From each generated output, we randomly sampled 60 sentences, 3 from each article for the sake of diversity. In this way, we obtained for each of the two languages 4 sets of 60 synthetic sentences (one from each model), plus 60 control sentences coming from the original human-written articles. After normalizing punctuation and mixing them randomly, we obtained a set of 300 sentences for each of the two languages. Each set is presented to a pool of 3 evaluators, who must decide for each sentence whether it has been created by a human or by Artificial Intelligence (AI). The only requirement for the evaluators is that they are native speakers of the language in question (Catalan and English, respectively). The final label is decided by a majority vote.

Finally, Table 1 shows the GPT-3 tokens per word[18] for each language for both Q&A and Text Summarization tasks, while Table 2 shows statistics of the summarization datasets. In the Appendix, we add additional statistics regarding the used datasets.

4. Results

In this section, we present the main results of the different experiments described above. We refer to the Appendix for a detailed description of the datasets and additional results.

Zero-shot Multilingual Question-Answering Figure 1 shows the F1 scores automatically obtained by the different GPT-3 models for each language in the Question-Answering task, by applying the evaluation script from SQuAD. Unsurprisingly, GPT-3 performs best in English, but the rest of the languages obtain remarkably good results, particularly taking into consideration the zero-shot scenario. In this task, we observe that the larger the model, the better the score, consistently for all languages.

Zero-shot Multilingual Text Summarization Figure 2 shows the results of the automatic evaluation of the Text Summarization task using ROUGE1[19] on the summaries obtained by the different models, compared to the reference. These results show two unanticipated results. One is that Catalan gets consistently better results than English, which only goes second. The other is that, contrary to the results obtained in the Q&A task (Figure 1) the performance decreases for the largest model, Davinci. Intrigued by the good results of Catalan and aware of the good quality of the CaSum training corpus (which was manually revised), we manually sampled 10% of the English corpus (i.e. 50 articles and summaries) and verified they were 100% correct.

As for the unexpected lower results of Davinci, a random manual inspection shows that Davinci tends to

[16] In the considered datasets, summaries are rarely longer than 3 sentences. We took the first 3 sentences to avoid taking unrelated sentences, since there is no marker for the end of the summary and we do not want to take e.g. the start of another article generated by GPT-3, in case the model follows the sequence `Article, TLDR, Summary, Article, TLDR, Summary,...` generating unrelated sentences from made-up articles.

[17] Note that we cannot just sample unconditionally since we need to force the model to write in the desired language.

[18] Tokens being GPT-3 tokens (subwords) and words being approximated with `wc` (i.e. with whitespaces).

[19] ROUGE-1 is a variant of the ROUGE metric that refers to the overlap of unigrams (i.e. each word) between the system and reference summaries

	CA	EN	ES	DE	TR
Article avg. #words	351.97	652.10	582.04	479.67	174.45
Summary avg. #words	32.82	54.07	21.49	23.61	25.85
Novelty	20.36	17.61	24.68	25.15	43.47
Compression ratio	11.83	13.01	29.67	21.68	7.08
Vocabulary size	29,365	42,265	50,160	49,502	30,384

Table 2: Statistics for the Summarization datasets. Novelty is the percentage of words in the summary that were not in the article. Compression ratio is the ratio between article and summary length (# words). Vocabulary size is the total number of unique words.

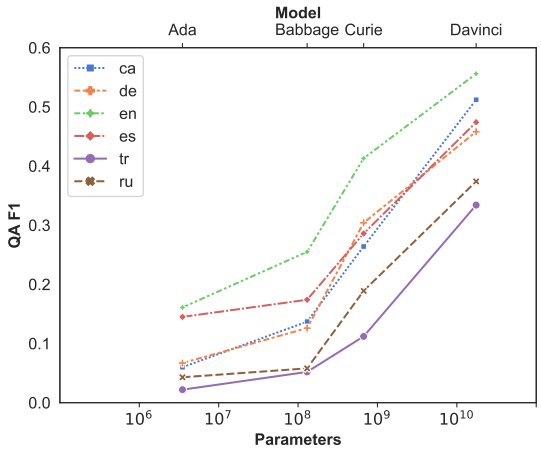


Figure 1: Automatic metrics results (F1) for the question-answering task

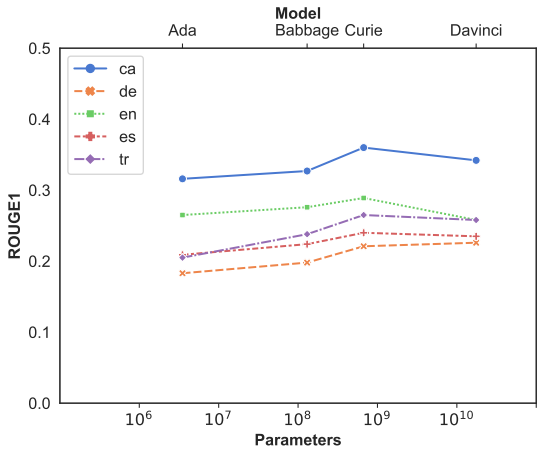


Figure 2: Automatic metrics results (ROUGE-1) for the Text Summarization task

yield more concise summaries and is more creative in terms of the lexical choices, using words that do not appear in the groundtruth.

Thinking that this may be unfairly penalising Davinci, we conduct an extra human evaluation for a subset of the English and Catalan summaries. We sample 75 articles, out of the 500 total, plus the respective 5 summaries (4 generated by the models and the reference) in English and in Catalan, and ask human evaluators to rank them. The resulting ranking is shown in Figure 3. Aside of a few obvious facts, such as that human summaries are ranked first more often (around half of the times), and that Ada (the smallest model) is ranked worst more often than the others, the rest of the results are less conclusive with respect to the effect of scale.

Multilingual Text Generation Figure 4 shows the results of the human evaluation of the sentences generated by the four models for Catalan and English. For each model, we report the labels (Human or AI) obtained through the majority vote of the 3 evaluators for each sentence. Considering that being mistaken with Human may be considered success for an AI model, we see that for English even the smaller models show an acceptable performance, and that Davinci is considered even "more human" than the human reference. As for Catalan, the results show a perfect correlation between size of the model and performance, with Davinci reaching a remarkable result, close to the one for English.

The interannotator agreement obtained per language in terms of Fleiss K is of 0.401 for Catalan and 0.290 for English, which are considered moderate and fair agreement, respectively.

5. Discussion

In this paper we are putting to test the usability of GPT-3 (trained on a humongous English corpus mixed with anecdotal amounts of data from an array of other languages) in multilingual scenarios. The initial hypothesis was that GPT-3 would yield acceptable results, close to the ones it obtains in English, in different generative tasks for a diverse set of languages. Overall, we can safely assume that the results obtained in the different experiments confirm our initial hypothesis.

In the Question-Answering task, the results may be considered very good for a zero-shot setting, when compared with supervised or semi-supervised systems.[20] In this experiment, the most remarkable finding is that scaling up the complexity of the model, we find that the gap between the results for English and the other languages tends to diminish in a way which is inversely proportional to the amount of data present for

[20] See e.g. mBERT baselines reported in Artetxe et al. (2019)

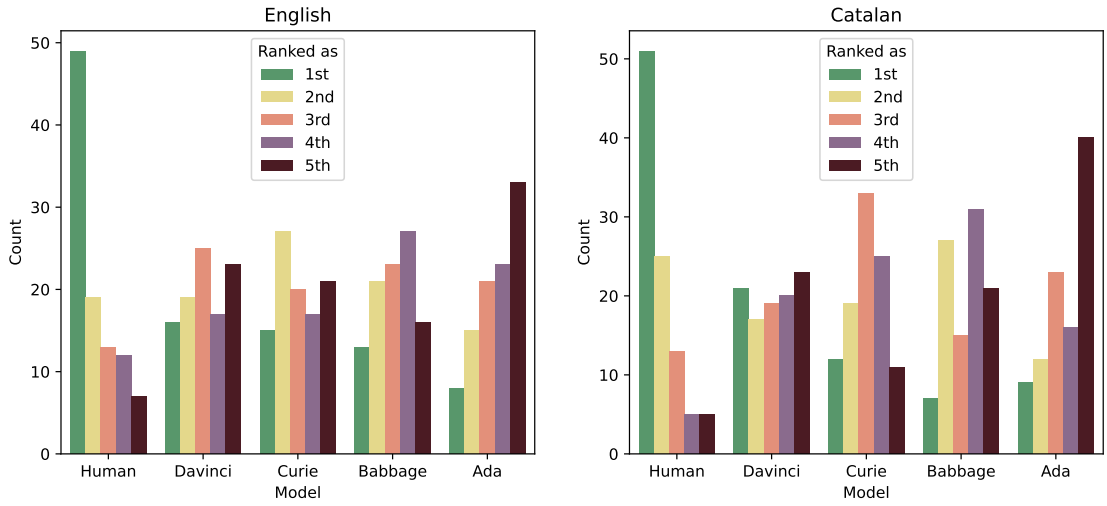


Figure 3: Human ranking results for the Text Summarization Evaluation task.

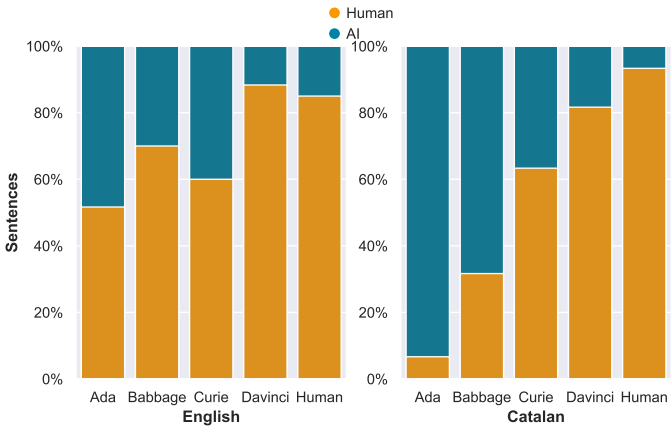


Figure 4: Human evaluation results for the Text Generation task

this language in the training corpus. More on this in the Scaling paragraph in Section 5.

Results in the Text Summarization task in terms of comparison between English and the other languages, not only confirm our initial hypothesis but unexpectedly show that GPT-3 is better at the task for Catalan than for English. Aware of the good quality of the CaSum dataset, we perform a further human review of a 10% sample of the English dataset (which had already been automatically filtered for quality using an existing summarization model and discarding summaries with low scores). The manual revision of the sample confirms also the quality of the English dataset, yielding a 100% of correct summaries. Since the datasets to evaluate this tasks are not parallel (contrary to Q&A), we may need to attribute all the discrepancies on the different degree of complexity of the respective datasets.

Table 2 provides the average length and other statistics for the five summarization datasets. Average length of the articles may be used as a proxy for the relative complexity of the task. English articles are the longest on average. In particular, they are almost twice as long as Catalan articles. In contrast, a higher rate of novelty may imply less overlapping with the groundtruth, i.e. being scored lower. Catalan, the best scored, has shorter articles and relative low novelty. English, which has longer articles but also low novelty, scores second. Very high novelty in Turkish summaries may be a reason for the low score, in spite of the articles being short on average. Meanwhile, Spanish and German present similar statistics and results. Small vocabulary size in the case of Catalan may also have an influence in the easiness of the task.

In our experimental setting for evaluating "pure" text generation we have tried to account for the intrinsic subjectivity of this task, by using majority vote in a system of three human evaluators. The task itself is completely subjective and requires that the evaluators have native language proficiency, able to capture subtleties. As a result we see that Davinci, with 88% of its sentences considered Human passes the Turing test with flying colors for English, especially considering

that only 85% of real human sentences in English pass as human. Importantly for our purposes, Davinci's performance in Catalan is almost as good, with a 81,6% of "human-looking" sentences. Not only that, Curie is as good in Catalan as it is in English, or actually better (63,3% and 60% respectively). It is true that Curie performance in English is atypical (in that it is lower than Babbage) and does not follow the expected effects of model scaling. More on that below.

Tokenization One of the simplest, yet crucial, aspects for understanding the multilingual performance of a language model is tokenization. In this work, the multilingual scenario is conditioned by the English-based vocabulary of GPT-3, made of the subword segmentation of English words. Table 1 presents the token/word ratio for each language. As our experiments show, this ratio turns out to be a useful predictor of GPT-3 performance for a given language. In the case of Russian, the Summarization task is not even possible because the Russian tokenization is such that average-length articles do not fit in GPT-3's context window of 2048 tokens. If we look at the results of the XQuAD experiment, languages can be clustered in 3 groups that are the same groups that emerge from Table 1: 1. English, with the best performance and fewer tokens per word; 2. Catalan, Spanish and German in the middle; and 3. Russian and Turkish with worse performance and more tokens per word.

Scaling The study of how scaling affects multilingual performance allows to forecast multilingual performance of future models. For instance, in the case of Q&A, in Figure 1, Catalan seems to be clearly closing the gap with English and should reach English levels in an hypothetical GPT-4. Interestingly, we could also predict at which point a zero-shot GPT would meet the current supervised SOTA models. As shown in the same figure, there is a steep curve of F1 score in terms of model size, while pre-training data (and, thus, the amount of non-English data) remains the same. This shows that transfer learning between English and the other languages in zero-shot settings scales with model size in a very steep curve. This is coherent with Figure H.11 in Brown et al. (2020), where zero-shot translation in which English is the target language reaches a plateau, but when the target languages are not English, the curves keep climbing.

However, we cannot fit proper scaling curves due to the lack of enough data points. Note that unlike in e.g. Kaplan et al. (2020), here we study downstream metrics, not loss. Downstream metrics should be more informative, but less smooth and more difficult to model than losses. In the case of Text Summarization, we have found a pattern clearly against what one would expect regarding scaling, that we attribute to the heterogeneity of the summarization datasets and the difficulties of evaluating this task. Finally, in the case of Text Generation, while the evaluated non-English language, Catalan, clearly improves performance along with increased model size, in the case of English we observe a less linear progression, with the Curie model seeming to stall. However, Davinci, the largest model does perform at top level. We may attribute this difference in the progression of the models to the fact that smaller GPTs are already quite proficient at generating good quality English sentences.

Usability in practice What we have found is that GPT-3 can be useful in multilingual applications, at least, in a degree not far from the one for English, especially considering that we use the model in zero-shot settings, yet still far from the supervised baselines. We expect the model to perform considerably better in few-shot settings (as opposed to zero-shot). We have seen that model scaling and English data compensate the lack of multilingual data in some scenarios, such as Q&A. And that in other scenarios, such as Summarization, smaller models can already give useful results. We have also seen that, currently, using GPT-3 (English-based) tokenization makes other languages expensive to compute, or even impossible (as Russian, for the MLSUM evaluation). A viable alternative would be to use multilingual tokenizers.

Limitations of our study In this study we have focused on five languages plus English and 3 generative tasks, but we believe that the results can be extrapolated to many more languages and other tasks. The main limitations of our study come from 1. the intrinsic difficulty of evaluating natural language generation, and 2. the lack of control over the models we are studying, which limits us to the 4 model sizes offered by the OpenAI API and prevents us from studying the effect of different tokenizations or data regimes. Nevertheless, we believe our work sets the grounds for future research in the intersection between multilingual language modeling, scaling, and cross-lingual transfer.

6. Conclusions and Future Work

We have seen that GPT-3 does, indeed, exhibit remarkable zero-shot generative capabilities in languages other than English that appear in tiny proportions in the training corpus, even for languages with non-Latin alphabets, like Russian, or with no typological affiliation, like Turkish. The results obtained on the different language evaluations are surprisingly close to the reference results for English for all tasks. This seems to confirm the extraordinary capacity of massive language models (even those mostly monolingual, like GPT-3) to generalise not only across tasks but most notably across languages, constituting an interlingua of sorts.

In general, our results show that right now GPT-3 can be almost as useful for many languages as it is for English, with room for improvement if optimization of the tokenization is addressed. On the overall, this is a very interesting exercise of how linguistic structures (universals) transfer across languages. Given the large amount of tasks GPT-3 has been implicitly exposed to

during the training procedure, handling a different language can be considered as working on yet another domain.

As future work, we suggest extending the study of the scaling laws of language models (Kaplan et al., 2020) in terms of cross-lingual transfer, similarly to Hernandez et al. (2021).

7. Acknowledgements

This work was funded by the MT4All CEF project.[21]

8. Bibliographical References

Armengol-Estapé, J., Carrino, C. P., Rodriguez-Penagos, C., de Gibert Bonet, O., Armentano-Oller, C., Gonzalez-Agirre, A., Melero, M., and Villegas, M. (2021). Are multilingual models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4933–4946, Online, August. Association for Computational Linguistics.

Artetxe, M., Ruder, S., and Yogatama, D. (2019). On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*.

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. *CoRR*, abs/2005.14165.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Dale, R. (2021). Gpt-3: What's it good for? *Natural Language Engineering*, 27(1):113–118.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Elkins, K. and Chun, J. (2020). Can gpt-3 pass a writer's turing test. *Journal of Cultural Analytics*, 2371:4549.

Floridi, L. and Chiriatti, M. (2020). Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694.

Henighan, T., Kaplan, J., Katz, M., Chen, M., Hesse, C., Jackson, J., Jun, H., Brown, T. B., Dhariwal, P., Gray, S., Hallacy, C., Mann, B., Radford, A., Ramesh, A., Ryder, N., Ziegler, D. M., Schulman, J., Amodei, D., and McCandlish, S. (2020). Scaling laws for autoregressive generative modeling. *CoRR*, abs/2010.14701.

Hernandez, D., Kaplan, J., Henighan, T., and McCandlish, S. (2021). Scaling laws for transfer. *CoRR*, abs/2102.01293.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. *CoRR*, abs/2001.08361.

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

Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., and Chen, W. (2021). What makes good in-context examples for gpt-3? *CoRR*, abs/2101.06804.

McGuffie, K. and Newhouse, A. (2020). The radicalization risks of gpt-3 and advanced neural language models. *arXiv preprint arXiv:2009.06807*.

Nikolich, A. and Puchkova, A. (2021). Fine-tuning gpt-3 for russian text summarization. *arXiv preprint arXiv:2108.03502*.

Nozza, D., Bianchi, F., and Hovy, D. (2020). What the [mask]? making sense of language-specific BERT models. *CoRR*, abs/2003.02912.

Radford, A. and Narasimhan, K. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

Sai, A. B., Mohankumar, A. K., and Khapra, M. M. (2020). A survey of evaluation metrics used for NLG systems. *CoRR*, abs/2008.12009.

Schluter, N. (2017). The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45, Valencia, Spain, April. Association for Computational Linguistics.

Scialom, T., Dray, P.-A., Lamprier, S., Piwowarski, B., and Staiano, J. (2020). Mlsum: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Zhao, T. Z., Wallace, E., Feng, S., Klein, D., and Singh, S. (2021). Calibrate before use: Improving

[21] https://ec.europa.eu/inea/en/connecting-europe-facility/cef-telecom/2019-eu-ia-0031

few-shot performance of language models. *arXiv preprint arXiv:2102.09690*.

9. Language Resource References

Artetxe, M., Ruder, S., and Yogatama, D. (2019). On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*.

Hermann, K. M., Kociský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Scialom, T., Dray, P.-A., Lamprier, S., Piwowarski, B., and Staiano, J. (2020). Mlsum: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067.

A. Model outputs and code

We publish both the outputs and the used code with an open license, with special emphasis on the model outputs, which we plan to release as a new dataset for analyzing multilingual skills of English GPT models.[22]

B. Question Answering additional results

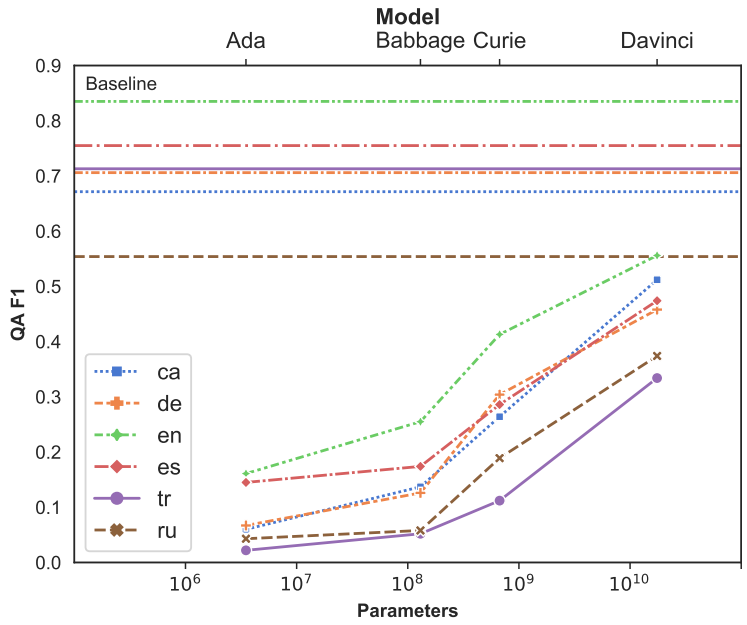


Figure 5: Automatic results (F1) for question-answering with the corresponding baselines

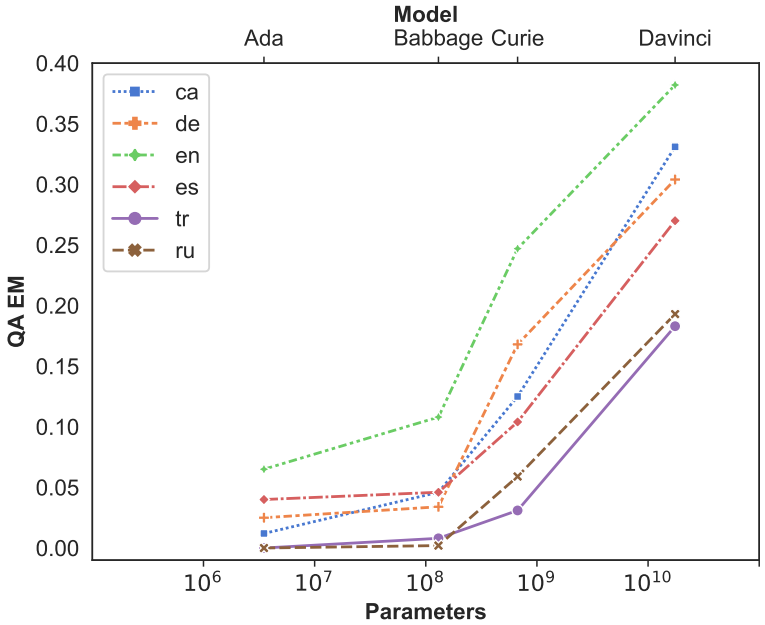


Figure 6: Automatic results (Exact Match) for question-answering

[22] https://github.com/TeMU-BSC/gpt3-queries

	CA		DE		EN		ES		RU		TR		avg.	
Model	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM
Ada	0.06	0.01	0.07	0.03	0.16	0.06	0.14	0.04	0.04	0.00	0.02	0.00	0.08	0.02
Babbage	0.14	0.05	0.13	0.03	0.25	0.11	0.17	0.05	0.06	0.00	0.05	0.01	0.13	0.04
Curie	0.26	0.13	0.30	0.17	0.41	0.25	0.29	0.10	0.19	0.06	0.11	0.03	0.26	0.12
Davinci	0.51	0.33	0.46	0.30	0.56	0.38	0.47	0.27	0.37	0.19	0.33	0.18	0.45	0.28

Table 3: Question-answering results reported as F1 and EM

	CA		DE		EN		ES		RU		TR		avg.	
Model	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM
Ada	0.06	0.01	0.07	0.03	0.16	0.06	0.14	0.04	0.04	0.00	0.02	0.00	0.08	0.02
Babbage	0.14	0.05	0.13	0.03	0.26	0.11	0.18	0.05	0.07	0.00	0.05	0.01	0.14	0.04
Curie	0.27	0.13	0.31	0.17	0.42	0.25	0.29	0.11	0.22	0.07	0.11	0.03	0.27	0.12
Davinci	0.52	0.33	0.47	0.31	0.56	0.39	0.48	0.27	0.44	0.23	0.33	0.18	0.47	0.29

Table 4: Question-answering results reported as F1 and EM with lemmatization

	CA		DE		EN		ES		RU		TR		avg.	
Model	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM
mBERT	0.67	0.47	0.71	0.54	0.84	0.72	0.76	0.57	0.71	0.53	0.55	0.40	0.71	0.54

Table 5: mBERT baseline reported as F1 and EM for Q&A, in (Artetxe et al., 2019) and (Armengol-Estapé et al., 2021) for Catalan

C. Summarization additional results

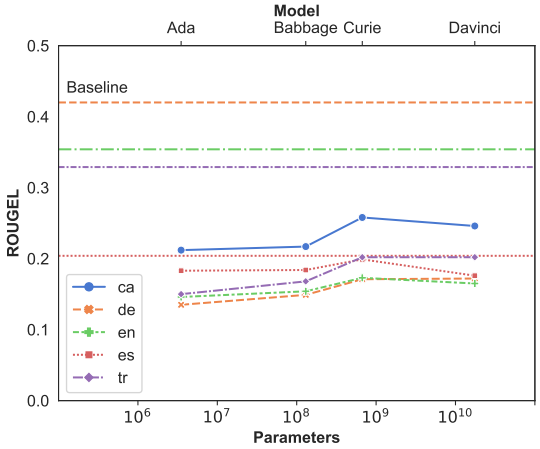
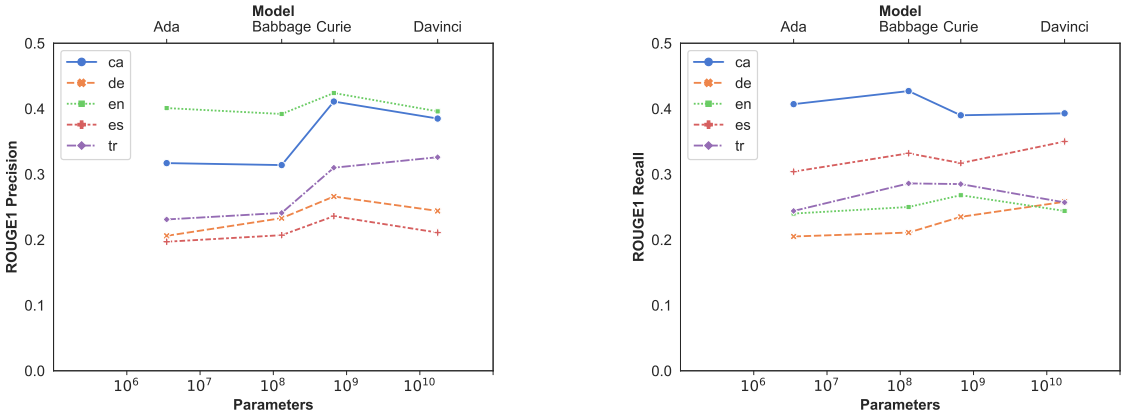


Figure 7: Automatic results (ROUGE-L) for summarization with the corresponding baselines



(a) Precision

(b) Recall

Figure 8: Automatic results (ROUGE-1 precision and recall) for summarization

	CA			DE			EN			ES			TR			avg.		
Model	R1	RL	M	R1	RL	M	R1	RL	M	R1	RL	M	R1	RL	M	R1	RL	M
Ada	0.32	0.21	0.27	0.18	0.14	0.15	0.27	0.18	0.19	0.21	0.15	0.17	0.21	0.15	0.14	0.24	0.17	0.18
Babbage	0.33	0.22	0.28	0.20	0.15	0.17	0.28	0.18	0.20	0.22	0.15	0.18	0.24	0.17	0.17	0.25	0.17	0.20
Curie	0.36	0.26	0.28	0.22	0.17	0.18	0.29	0.20	0.21	0.24	0.17	0.19	0.27	0.20	0.18	0.28	0.20	0.21
Davinci	0.34	0.25	0.28	0.23	0.17	0.20	0.26	0.18	0.19	0.24	0.16	0.21	0.26	0.20	0.17	0.26	0.19	0.21

Table 6: Summaritzation results reported as ROUGE-1 (R1), ROUGE-L (RL) and METEOR (M)

	CA		DE		EN		ES		RU		TR		avg.	
Model	RL	M	RL	M	RL	M	RL	M	RL	M	RL	M	RL	M
mBERT	-	-	0.42	0.26	0.35	0.22	0.20	0.15	0.09	0.07	0.33	0.263	0.33	0.22

Table 7: mBERT baseline reported as ROUGE-L (RL) and METEOR (M) for summarization from (Scialom et al., 2020)

C.1. CaSum: the Catalan Summarization dataset

CaSum is a summarization dataset extracted from a newswire corpus crawled from the Catalan News Agency.[23] The corpus consists of a collection of 217,735 articles together with their summary. The summaries have been automatically constructed by joining the original headline and subtitle of each article. This is a usual technique to automatically build summarization corpora, common to most of MLSUM datasets.

C.2. Summarization models used for sampling

- German: `T-Systems-onsite/mt5-small-sum-de-en-v2`
- English: `T-Systems-onsite/mt5-small-sum-de-en-v2`
- Spanish: `Narrativa/bsc_roberta2roberta_shared-spanish-finetuned-mlsum-summarization`
- Turkish: `mrm8488/bert2bert_shared-turkish-summarization`

D. Text Generation additional results

	ENGLISH				CATALAN			
	HUMAN		AI		HUMAN		AI	
Ada	31	51.67%	29	48.33%	4	6.67%	56	93.33%
Babbage	42	70.00%	18	30.00%	19	31.67%	41	68.33%
Curie	36	60.00%	24	40.00%	38	63.33%	22	36.67%
Davinci	53	88.33%	7	11.67%	49	81.67%	11	18.33%
Human	51	85.00%	9	15.00%	56	93.33%	4	6.67%

Table 8: Human Evaluation results for text generation

[23] `https://www.acn.cat/`