High-quality Data-to-Text Generation for Severely Under-Resourced Languages with Out-of-the-box Large Language Models

Michela Lorandi and **Anya Belz**
ADAPT Research Centre, Dublin City University
{michela.lorandi, anya.belz}@adaptcentre.ie

Abstract

The performance of NLP methods for severely under-resourced languages cannot currently hope to match the state of the art in NLP methods for well resourced languages. We explore the extent to which pretrained large language models (LLMs) can bridge this gap, via the example of data-to-text generation for Irish, Welsh, Breton and Maltese. We test LLMs on these under-resourced languages and English, in a range of scenarios. We find that LLMs easily set the state of the art for the under-resourced languages by substantial margins, as measured by both automatic and human evaluations. For all our languages, human evaluation shows on-a-par performance with humans for our best systems, but BLEU scores collapse compared to English, casting doubt on the metric's suitability for evaluating non-task-specific systems. Overall, our results demonstrate the great potential of LLMs to bridge the performance gap for under-resourced languages.

1 Introduction

Automatically generating text for a given data set (e.g. a textual summary) is a much bigger challenge for severely under-resourced languages than for well resourced languages like English. Creating a rule-based system by hand is one option: slow but faster if language-independent resources can be used (Mille et al., 2023). An alternative is task-specific finetuning and collecting training data for it (partly) by hand and/or by collecting/generating silver training data which may be good enough to achieve a desired performance level.

These methods all take varying but considerable amounts of manual work and time. In contrast, using large language models (LLMs) in their 'out of the box' state has next to no such overheads. However, at this point their zero-shot ability to generate correct text of sufficient quality (e.g. in terms of minimum real-world usefulness where first-draft plus post-editing takes less time than from-scratch) for severely under-resourced languages is untested.

Given that by definition LLMs will have seen very little text in under-resourced languages during training, using them in zero-shot mode for text generation in such languages may not seem a promising idea. In this paper, we explore the extent to which it is possible for data-to-text generation, in so doing shedding light on the potential of LLMs to bridge performance gaps between under-resourced languages (the vast majority of the world's languages) and well resourced languages like English.

All code and results are available on GitHub: https://github.com/michelalorandi/D2T-Gen-for-Under-Res-Lang-w-LLMs.

2 Related Research

A large number of papers in the past year have reported work on using LLMs, and GPT in particular, in zero or few-shot mode for a wide range of different tasks, including both system development (Liu et al., 2023; Long, 2023; Lu et al., 2022; Wang et al., 2023b; Qin et al., 2023) and evaluation (Chiang and Lee, 2023; Wang and Chang, 2022; Chan et al., 2023; Shen et al., 2023; Hada et al., 2023).

Because the performance of zero-shot LLMs depends on the quality of the prompt, there has been a corresponding flurry of research on prompt engineering, including plan-and-solve prompting (Wang et al., 2023a), tree-of-thought prompting (Yao et al., 2023; Long, 2023), and automatic prompt fixing (Pearce et al., 2023).

WebNLG 2023 (see below) included a first attempt (Lorandi and Belz, 2023) to perform data-to-text generation for under-resourced languages using out-of-the-box GPT-3.5 plus Google Translate which outperformed other participating systems by considerable margins. We take the same approach but test four LLMs and three MT systems (two closed source and one open source) in a wider range of scenarios, and additionally test our best

Findings of the Association for Computational Linguistics: EACL 2024, pages 1451–1461
March 17-22, 2024

system on English where the tough state-of-the-art outperforms humans.

3 Data and Task

WebNLG 2023 is the third iteration of the WebNLG shared task series and focuses on the severely under-resourced European languages Irish, Breton, Welsh and Maltese[1] (Cripwell et al., 2023). The WebNLG 2023 data consists of 1,778 test items for each language, 1,399 dev items for Breton, and 1,665 dev items for Welsh, Irish and Maltese. The test sets were manually translated by professional translators from the English originals. Additionally 13,211 training items are provided where texts were automatically translated from English.

WebNLG 2023 systems map from RDF triples to a suitable output text, as in the example from the WebNLG'23 website[2] in Figure 1. The complete shared-task data is available from the same website.

(a) Set of RDF triples

```
<entry category="Company" eid="Id21"
shape="(X (X) (X) (X) (X))"
shape_type="sibling" size="4">
  <modifiedtripleset>
    <mtriple>Trane | foundingDate |
    1913-01-01</mtriple>
    <mtriple>Trane | location | Ireland
    </mtriple>
    <mtriple>Trane | foundationPlace |
    La_Crosse,_Winsconsin</mtriple>
    <mtriple>Trane | numberOfEmployees
    | 29000</mtriple>
  </modifiedtripleset>
</entry>
```

(b) English text

Trane, which was founded on January 1st 1913 in La Crosse, Wisconsin, is based in Ireland. It has 29,000 employees.

Figure 1: WebNLG input set of triples and output text.

4 Models

We test four different pretrained LLMs (paid-for GPT-3.5, and open-source Bloom, LLaMa2-chat, and Falcon-chat), each in two modes: (i) direct generation into the target language, and (ii) generation into English followed by translation into the target language with one of three machine translation (MT) engines (Google Translate, Alibaba Translate, and No Language Left Behind system (Costa-jussà et al., 2022)).

[1]https://synalp.gitlabpages.inria.fr/webnlg-challenge/challenge_2023/

[2]https://synalp.gitlabpages.inria.fr/webnlg-challenge/docs

GPT-3.5 or InstructGPT (Ouyang et al., 2022) is GPT-3 plus supervision fine-tuning on instruction data, reward model training and Reinforcement Learning with Human Feedback (RLHF) with the reward model. **BLOOM** (Scao et al., 2022) is trained on the ROOTS corpus, a collection of 498 HuggingFace datasets. **LLaMa2-chat** (Touvron et al., 2023) builds on the pretrained LLaMa2 model (trained only on publicly available datasets) fine-tuned in two steps similar to GPT-3.5, but instead of using one reward model for helpfulness and safety, two separately optimised reward models are used. **Falcon-chat** (Almazrouei et al., 2023) builds on Falcon-base, which is trained on the RefinedWeb dataset (Penedo et al., 2023). Falcon-base is then fine-tuned on chat and instruction datasets with a mix of large-scale conversational datasets.

5 Experimental Set-up

In this section we describe the main aspects of the experimental set-up. Hyperparameters and API access are provided in Section A.1 in the appendix.

5.1 Experimental grid

We tested all combinations of our four LLMs, two translation engines, two prompts, and five languages, i.e. the basic experimental grid looks as follows: {GPT-3.5, Bloom, Llama2, Falcon} $\times$ {Google Translate, Alibaba Translate, NLLB system} $\times$ {zero-shot minimal instruction, few-shot in context} $\times$ {Irish, Breton, Maltese, Welsh, English}.

5.2 Prompt engineering

We use the prompts previously identified (Lorandi and Belz, 2023) as the most suitable for data-to-text generation following prompt testing of zero-shot minimal instruction, few-shot in-context learning, and chain-of-thought (CoT) (Wei et al., 2022) on GPT-3.5 and GPT-4, on a different random sample of 20 data/text pairs in each phase.

For the work reported here, we conducted a preliminary testing phase with BLOOM, LLaMa2, and Falcon to verify if further postprocessing is needed. As a result, we remove all Python code, occurrences of """, and output start markers (e.g. "Falcon:") from the output of all three.

5.3 Evaluation

We carried out automatic evaluations with BLEU (Papineni et al., 2002), ChrF++ (Popović, 2017)

M	Prompt	Irish			Welsh			Maltese			Breton		
		BLEU↑	**ChrF++↑**	**TER↓**	**BLEU↑**	**ChrF++↑**	**TER↓**	**BLEU↑**	**ChrF++↑**	**TER↓**	**BLEU↑**	**ChrF++↑**	**TER↓**
GPT-3.5 (175B)	ZS MI	12.9931	0.4124	0.9298	15.8695	0.4619	0.822	13.0311	0.445	0.8496	16.4171	0.4303	0.7813
	FS IC	15.3477	0.4303	0.8451	18.9512	0.4742	0.7192	15.4315	0.4536	0.7605	***18.5925***	***0.4473***	***0.7218***
	ZS MI +GT	***20.5176***	***0.5146***	0.7122	24.7126	***0.5496***	0.6659	20.3528	***0.5263***	0.67	-	-	-
	FS IC + GT	20.4001	0.51	***0.6894***	***25.115***	0.5484	***0.6435***	***21.2656***	0.5249	***0.6465***	-	-	-
	ZS MI + AT	18.3807	0.4984	0.7184	23.4782	0.5408	0.6724	16.8312	0.4902	0.72	10.5379	0.3558	0.7954
	FS IC + AT	18.3433	0.495	0.6987	23.8908	0.5412	0.6493	17.5723	0.4867	0.6935	10.2411	0.3501	0.7864
	ZS+NLLB	17.5042	0.455	0.7356	19.294	0.4761	0.6948	16.457	0.4811	0.7262	-	-	-
	FS+NLLB	17.1448	0.4503	0.7136	19.106	0.4718	0.6782	17.1262	0.479	0.7015	-	-	-
BLOOM (176B)	ZS MI	2.6099	0.2118	2.8781	1.8576	0.2043	3.0441	2.7287	0.2303	2.9191	1.1293	0.161	1.8799
	FS IC	4.9828	0.2535	1.5027	6.558	0.2696	1.1825	9.4622	0.3075	*0.9589*	5.6066	0.2585	0.9923
	ZS MI +GT	6.6329	0.3672	2.2041	7.4595	0.3882	2.1584	6.3703	0.3745	2.0717	-	-	-
	FS IC + GT	*14.8148*	*0.4521*	*0.9073*	*15.4467*	*0.4683*	*0.9699*	*12.7663*	*0.4498*	0.9685	-	-	-
	ZS MI + AT	6.2173	0.36	2.1451	7.3117	0.3846	2.1301	5.6348	0.3552	2.1202	4.5007	0.2808	1.1941
	FS IC + AT	12.2466	0.4309	1.018	14.8386	0.4621	0.9889	10.7619	0.4229	1.0116	*8.2509*	*0.3197*	*0.8768*
	ZS+NLLB	4.9851	0.2563	1.4959	5.6246	0.2589	1.5071	4.8973	0.2607	1.2322	-	-	-
	FS+NLLB	7.6891	0.2708	1.0133	8.5701	0.2701	1.0038	6.4824	0.2705	0.9173	-	-	-
LLaMa2-chat (70B)	ZS MI	6.4367	0.2349	1.2706	6.6383	0.2529	1.1016	10.3055	0.3198	0.8965	4.0113	0.2147	0.8731
	FS IC	10.4064	0.364	1.0677	8.1874	0.3344	1.3614	12.5935	0.3901	0.8266	*10.2303*	0.3286	0.8095
	ZS MI +GT	16.7841	0.4872	0.8366	19.8404	0.5212	0.8052	16.7342	0.5028	0.7861	-	-	-
	FS IC + GT	*19.3366*	*0.5033*	*0.7378*	*23.6408*	*0.5412*	*0.6969*	*19.7145*	*0.5186*	*0.6903*	-	-	-
	ZS MI + AT	16.0344	0.4772	0.8391	19.3043	0.5139	0.8124	13.7873	0.471	0.8354	9.559	0.3438	0.8448
	FS IC + AT	17.9225	0.4907	0.7458	22.5067	0.5318	0.706	15.6232	0.4786	0.7478	10.0142	*0.3492*	*0.8007*
	ZS+NLLB	15.1903	0.4195	0.8259	16.8335	0.4429	0.7988	14.5649	0.4542	0.8111	-	-	-
	FS+NLLB	16.5713	0.442	0.7549	18.3623	0.4632	0.7208	15.7702	0.4648	0.7392	-	-	-
Falcon-chat (180B)	ZS MI	6.3239	0.2703	1.3245	6.0496	0.2679	1.4255	6.793	0.2765	1.3012	7.9701	0.2638	0.923
	FS IC	11.2338	0.3657	0.9902	13.0723	0.3611	0.8821	12.2097	0.3656	0.8725	9.749	0.3221	0.8079
	ZS MI +GT	13.4874	0.4584	1.1768	15.4119	0.486	1.1724	12.9136	0.467	1.1015	-	-	-
	FS IC + GT	*19.6085*	*0.5034*	*0.7453*	*23.1749*	*0.5387*	*0.7124*	*19.5894*	*0.5158*	*0.6907*	-	-	-
	ZS MI + AT	12.5954	0.4496	1.176	14.7283	0.4803	1.1743	10.6168	0.4379	1.1574	8.5235	0.3345	0.8977
	FS IC + AT	17.4847	0.4916	0.7536	22.5094	0.5327	0.7152	15.9008	0.4793	0.7486	*10.285*	*0.3503*	*0.8006*
	ZS+NLLB	12.9335	0.4012	1.1023	13.8666	0.4249	1.0798	11.2754	0.4253	1.074	-	-	-
	FS+NLLB	16.1999	0.4385	0.7573	18.5609	0.4631	0.7238	15.4012	0.4623	0.74	-	-	-
WebNLG23	FORGe	16.66	0.44	0.75	-	-	-	-	-	-	-	-	-
	IREL	-	-	-	20.97	0.49	0.67	16.49	0.47	0.7	-	-	-
	CUNI-Wue	-	-	-	-	-	-	-	-	-	10.09	0.33	0.80
	Baseline	11.63	0.36	0.74	10.70	0.36	0.77	15.60	0.42	0.67	9.92	0.33	0.76

Table 1: Automatic evaluation results for **Irish**, **Welsh**, **Maltese** and **Breton**. Highest score in each column for each language in bold, highest score for each model in italics. Number of parameters in brackets in column 1. ZS MI=Zero-Shot Minimal Instruction, FS IC=Few-Shot In Context, GT=Google Translate, AT=Alibaba Translate, NLLB=No Language Left Behind system.

and TER (Snover et al., 2006) for all systems (each cell in the experimental grid from Section 5.1); the resulting scores are shown in Table 1. Furthermore, we computed COMET (Rei et al., 2020) for all systems, and BERTScore (Zhang et al., 2019) for all Irish, Welsh and Breton systems (see Appendix B).

We report a new human evaluation of four of the English systems using exactly the same method as in WebNLG 2023 (Cripwell et al., 2023). In terms of the experimental grid above, the four systems in the human evaluation were {GPT-3.5} × {} × {zero-shot minimal instruction, few-shot in context} × {English}. We evaluated these alongside the best English system from WebNLG 2020, and the human-authored test-set outputs.

We also include relevant results from the WebNLG 2020 and 2023 human evaluations, from the latter for {GPT-3.5} × {Google Translate} × {few-shot in context} × {Irish, Maltese, Welsh}, and the second best WebNLG 2023 system.

6 Results

This section reports the main human and metric evaluation results. Details of cost in Section A.2.

6.1 Metrics

Metric results (BLEU, ChrF++ and TER) for all systems in our grid from Section 5.1 are shown in Table 1 for Irish/Welsh/ Maltese/Breton, and in Table 2 for English. Tables 6, 7 and 8 present BERTScore and COMET metric results for Irish/Welsh/Breton, English, and Irish/Welsh/Maltese/Breton/English, respectively.

High-level results across all languages are that GPT-3.5+GoogleTrans always has a higher metric score than all other model/translation engine combinations, except for English where it has the highest score for ChrF++, but is outperformed by the top-ranking WebNLG 2020 system for BLEU and TER.

Generation into English plus Google Translate has better scores than direct generation into the under-resourced language by substantial margins in all cases. Alibaba has slightly better scores than direct generation in all cases except Breton, while NLLB has slightly better scores than direct generation, but worse than Alibaba, in the majority of cases.

For all models except GPT, the few-shot version of a system is always better than the zero-shot. For GPT the few-shot and zero-shot results are much closer, and in a few cases, zero-shot is slightly better than few shot, e.g. for Maltese using translation.

For the under-resourced languages, the overall best metric scores are obtained for Welsh, by good margins, followed by Maltese, Irish, and Breton, where we cannot use Google Translate, and where in fact generation into English plus Alibaba is a lot worse than direct generation in case of GPT-3.5. This is in contrast to the other languages where Alibaba always achieves small improvements.

Considering COMET (Table 8), we get similar results for GPT-3.5 and Falcon-chat when using a MT system and Few-Shot In-Context prompt in all under-resourced languages.

An interesting aspect of the metric results is that while best BLEU scores are far higher for English than for any other language (e.g. more than twice as high for the best results), this pattern is not replicated in the ChrF++, TER, BERTScore and COMET scores. See Section 7 for discussion.

6.2 Human evaluation of English systems

Outside of WebNLG 2023, there is no state of the art for data-to-text generation in our four under-resourced languages that we can compare against.

Model	Prompt	BLEU ↑	ChrF++ ↑	TER ↓
GPT-3.5	ZS MI	49.6603	0.6895	0.4498
(175B)	FS IC	*52.7366*	***0.6906***	*0.42*
BLOOM	ZS MI	13.4535	0.4572	0.705
(176B)	FS IC	*32.1397*	*0.5816*	*0.5876*
LLaMa2-chat	ZS MI	40.4711	0.6421	0.5746
(70B)	FS IC	*46.8566*	*0.6705*	*0.4853*
Falcon-chat	ZS MI	31.3463	0.5922	0.6545
(180B)	FS IC	*46.3762*	*0.668*	*0.4891*
WebNLG 2020:				
Baseline FORGE2020		40.6	62.1	51.7
Amazon AI (Shanghai)		**54.0**	69.0	**40.6**
OSU Neural NLG		53.5	68.8	41.6

Table 2: Automatic evaluation results for **English**. Best score per column in bold, best score per model in italics. Number of model parameters in brackets. ZS MI=Zero-Shot Minimal Instruction, FS IC=Few-Shot In Context.

However, we can compare our methods against the best performing systems in English from WebNLG 2020, and we did this using the same human evaluation approach that was used in WebNLG 2023.

Table 3 shows the results from this evaluation of Fluency, Absence of Additions, and Absence of Omissions which show that few-shot GPT3.5 has the highest mean score for Fluency, Omissions and Repetition, with zero-shot having the highest mean in Additions. However, there are significant performance differences only for Omissions, reflecting a similar relatively lower score for Omissions in the WebNLG20 evaluations (see next section).

System	Fluency		Addition		Omission		
GPT-3.5 FS MI	**4.50**	A	0.88	A	**0.93**	A	
Amazon AI	4.33	A	0.90	A	0.82		B
GPT-3.5 ZS IC	4.33	A	**0.91**	A	**0.93**	A	
Human ref	4.28	A	0.83	A	0.92	A	B

Table 3: Human evaluation results for **English** for human-authored references, GPT-3.5 zero-shot, GPT-3.5 few-shot), and best WebNLG20 system. Means and homogeneous subsets from Tukey HSD (alpha = .05).

6.3 WebNLG human evaluations

Table 4 shows mean **WebNLG 2023** human scores for **Welsh**, **Maltese** and **Irish**, per system for Fluency, Addition and Omission, for the human reference texts, the GPT-3.5+Google Translate+few-shot system (DCU-NLG-PBN) and the next best system.

Here too, the differences between the scores for the human references and the DCU-NLG-PGN system (few-shot GPT + GT) are not statistically sig-

L	System	Fluency			Addition		Omission	
Welsh	Human ref	**3.28**	A		**0.9**	A	**0.84**	A
	DCU-NLG-PBN	3.25	A		0.86	A	0.77	A
	IREL	2.67	B		0.6	B	0.47	B
Maltese	Human ref	**4.27**	A		0.89	A	0.85	A
	DCU-NLG-PBN	4.06	A B		**0.91**	A	**0.86**	A
	IREL	3.74	B		0.69	B	0.56	B
Irish	Human ref	**4.07**	A		0.81	A	0.82	A
	DCU-NLG-PBN	3.83	A B		0.83	A	**0.85**	A
	DCU/TCD-FORGe	3.35		C	**0.84**	A	0.81	A

Table 4: Mean **WebNLG 2023** human scores for **Welsh**, **Maltese** and **Irish**, per system for Fluency, Addition and Omission.

nificant for any of the nine sets of scores; the human references come top 5 times, DCU-NLG-PGN 3 times, and DCU/TCD-FORGe once. The human references and the DCU-NLG-PBN system are significantly better than the runner up system for Maltese and Welsh on all evaluation criteria. Taken together, we can consider that on-par-with-human performance for the GPT+MT systems.

In Table 5, we show results for the **English** human evaluation from **WebNLG 2020** for reference (evaluation criteria translated to match our terminology).

L	System	Fluency		Addition		Omission	
English	Amazon AI	**90.286**	A	**95.196**	A	94.393	A
	OSU Neural NLG	90.066	A	94.615	A	95.123	A
	Human ref	89.846	A	94.392	A	**95.442**	A

Table 5: Human evaluation results of **English** from **WebNLG 2020**.

The two systems have slightly higher scores than the human references except for Omissions. Recall that Table 3 indicates that GPT3.5+MT outperforms the Amazon AI system and the human references. Taken together the two human evaluations indicate overall better performance for GPT3.5+MT.

7 Discussion and Conclusion

One striking aspect of the metric results for the under-resourced languages is that BLEU scores are far lower across the board than for English. At the same time, human evaluations show on-a-par-with-human performance for both the under-resourced languages and English. This shows a significant performance failure for BLEU that is not reflected in ChrF++, TER, BERTScore or COMET.

This BLEU failure may be due to two aspects: for one, BLEU is a word n-gram overlap metric, while ChrF++ and TER are character F-Score and character edit distance based, respectively. BERTScore computes cosine similarity for each token in candidate and reference sentences using the pre-trained contextual embeddings from BERT, and COMET uses a pretrained multilingual model trained to mimic human judgement. Two, the GPT training data is likely to have contained the English WebNLG data in its entirety (albeit not as input/output pairs), but not any of the under-resourced language outputs. It seems that under these circumstances, where system outputs and reference texts have not been sampled from the same narrow distribution, BLEU simply does not work.

The systems that we introduce and test here are generic, non-task-specifically trained systems. All of the systems we compare them against are task-specifically supervision-trained systems, and in one case (Mille et al., 2023), hand-crafted to perform a single specific task. It is yet another piece of evidence showcasing the astonishing out-of-the-box abilities of the latest generation of LLMs. Similarly to previous evidence, we see that absence of instruction tuning (BLOOM) and smaller size (LLaMa2) are associated with poorer performance. It is also unclear how such systems can be utilised in real-world application scenarios. However, we show the incredible ability of LLMs to generate texts on-a-par performance with humans for our best systems in all languages tested.

Limitations

In this work, we focused on the usage of LLMs together with MT engines. Not all the models used are open-sourced and to access them we need to use paid APIs. This not only implies a financial cost that could be prohibited, but also implies problems in terms of reproducibility as we're not entirely sure of what the model is behind the APIs.

Furthermore, considering the open-sourced LLMs, we need a large number of GPUs to be able to execute such models, especially BLOOM (176B) and Falcon (180B). In the case of Falcon, we would need at least 400GB of memory to run the model in inference.

Lastly, we explored only two simple types of prompts designed based on GPT-3.5 and it could be beneficial to explore more advanced types of prompts also taking into account differences between models.

Ethics Statement

We focused on under-resourced languages setting a base for further research and the development of real-world applications that people who speak such languages could use. On the other hand, when using LLMs there is a general risk that they could produce offensive or incorrect content that may harm people using such systems. Since our approach only takes into account the given input without any factual checking, we cannot guarantee that there is no generation of factually incorrect texts.

Furthermore, it's currently unclear what has been included in the training data of some LLMs, meaning that there may be evidence of bias in generated texts, which in turn carries a risk of possibly causing harm to the end user.

Acknowledgements

The cost of accessing the GPT API was covered by financial support from the DCU-NLG Research Group at DCU. Michela Lorandi's work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224. Both authors benefit from being members of the ADAPT SFI Research Centre at Dublin City University, funded by the Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Maitha Alhammadi, Mazzotta Daniele, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of language models: Towards open frontier models.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Liam Cripwell, , Anya Belz, Claudia Borg, Claire Gardent, Albert Gatt, John Judge, Michela Lorandi, Anna Nikiforoskaya, William Soto-Martinez, and Craig Thomson. 2023. The 2023 webnlg shared task on low resource languages overview and evaluation results (webnlg 2023). In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge*, Prague, Czech Republic.

Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2023. Are large language model-based evaluators the solution to scaling up multilingual evaluation? *arXiv preprint arXiv:2309.07462*.

Zhengliang Liu, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Wei Liu, Dinggang Shen, Quanzheng Li, et al. 2023. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*.

Jieyi Long. 2023. Large language model guided tree-of-thought. *arXiv preprint arXiv:2305.08291*.

Michela Lorandi and Anya Belz. 2023. Data-to-text generation for severely under-resourced languages with GPT-3.5: A bit of help needed from Google Translate (WebNLG 2023). In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 80–86, Prague, Czech Republic. Association for Computational Linguistics.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Simon Mille, Elaine U'i Dhonnchadha, Stamatia Dasiopoulou, Lauren Cassidy, Brian Davis, and Anya Belz. 2023. DCU/TCD-FORGe at WebNLG'23: Irish rules! In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge*, Prague, Czech Republic.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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

Hammond Pearce, Benjamin Tan, Baleegh Ahmad, Ramesh Karri, and Brendan Dolan-Gavitt. 2023. Examining zero-shot vulnerability repair with large language models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 2339–2356. IEEE.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Chenhui Shen, Liying Cheng, Yang You, and Lidong Bing. 2023. Are large language models good evaluators for abstractive summarization? *arXiv preprint arXiv:2305.13091*.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*.

Yau-Shian Wang and Yingshan Chang. 2022. Toxicity detection with generative prompt-based inference. *arXiv preprint arXiv:2205.12390*.

Zihao Wang, Shaofei Cai, Anji Liu, Xiaojian Ma, and Yitao Liang. 2023b. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *arXiv preprint arXiv:2302.01560*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

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

A Appendix

A.1 Hyperparameters and APIs

We executed all the experiments either via API or on our own GPUs. We used the paid-for OpenAI API to access text-davinci-003 [3] (GPT-3.5), while we used the free inference API of HuggingFace to access BLOOM 176B [4] and falcon-180B-chat [5]. On the other hand, we downloaded and executed Llama-2-70b-chat-hf[6] on a Nvidia A100 GPU with 80GB RAM.

To use the three explored Machine Translation engines, we used the pay-as-you-go APIs of Google Cloud [7] and Alibaba Cloud [8], and we downloaded and executed NLLB (Costa-jussà et al., 2022) on a Nvidia A100 GPU with 80GB RAM.

For all used models, we set *maximum length* to 500 with Zero-Shot Minimal Instruction and 1000

[3] https://platform.openai.com/docs/models/gpt-3-5

[4] https://huggingface.co/bigscience/bloom

[5] https://huggingface.co/tiiuae/falcon-180B-chat

[6] https://huggingface.co/meta-llama/Llama-2-70b-chat-hf

[7] https://cloud.google.com/translate

[8] https://www.alibabacloud.com/product/machine-translation

with Few-Shot In Context. All generated texts are post-processed as described above.

GPT-3.5 In all experiments involving GPT-3.5, we set text-davinci-003 parameters to *temperature*=0, *top p*=1 (default), *frequency penalty*=0 and *presence penalty*=0 (default), *best of*=1 (default) to get only 1 completion for each prompt.

BLOOM We used bigscience/bloom model with HuggingFace's Inference Client API setting the parameters to *temperature*=0.7, *top p*=0.9, *frequency penalty*=0 and *presence penalty*=0.

LLaMa2-chat We used meta-llama/Llama-2-70b-chat-hf model on HuggingFace setting the parameters to *temperature*=1 (default), *top p*=1 (default), *repetition penalty*=1 (default) and *diversity penalty*=0 (default), *num return sequences*=1.

Falcon-chat We used tiiuae/falcon-180b-chat model with HuggingFace's Inference Client API setting the parameters to *temperature*=0.7, *top p*=0.9, *frequency penalty*=0 and *presence penalty*=0.

NLLB We used facebook/nllb-200-1.3B model on HuggingFace setting the languages to *mlt_Latn*, *cym_Latn*, and *gle_Latn*, respectively for Maltese, Welsh, and Irish.

COMET We used the Unbabel/wmt22-comet-da model on HuggingFace.

A.2 Computational and financial cost

To execute our experiments, we relied on the use of paid APIs and GPU usage.

Considering paid APIs, GPT-3.5 model cost US$91.82 in API, while the usage of Google Translate and Alibaba cost respectively €135.15 and US$377.97.

Regarding computational time and cost, we executed all LLama2 chat experiments on a Nvidia A100 GPU, which took, on average, around 21 hours to execute a single experiment using Zero-Shot Minimal Instruction (ZS MI) prompt and around 2 days and 18 hours to execute a single experiment using Few-Shot In Context (FS IC) prompt. On the other hand, we accessed all the other models through API calls. On average, using HuggingFace inference API BLOOM176B took around 17 hours for ZS MI prompt and around 2 days for FS CI prompt, while Falcon 180B took around 11 hours for ZS MI prompt and around 20 hours for FS CI prompt. Lastly, using GPT-3.5 with OpenAI APIs, it took around 1 hour both for ZS MI and FS CI prompts.

B Additional results

In this Section, we provide additional automatic evaluation results using BERTScore and COMET.

Tables 6 and 7 present BERTScore results for all systems in Irish/Welsh/Breton and English, respectively. Maltese is not included as it is not available in BERTScore.

Tables 8 present COMET results for all systems in our grid from Section 5.1, for Irish/Welsh/Maltese/Breton/English.

C Prompts

We provide the prompts we used to execute all our experiments. In Table 9 Zero-Shot Minimal Instruction prompt is shown, while in Table 10 Few-Shot In Context prompt is shown with the examples used for each language tested.

D Human evaluation setup

For our human evaluation of English systems, we considered the human-authored references, GPT-3.5 Zero-Shot Minimal Instruction prompt, GPT-3.5 Few-Shot In Context prompt, and the best WebNLG2020 system (Amazon AI). For each system, we annotated 100 samples recruiting 4 annotators, who are non-author members of the research group plus one close collaborator.

We followed the same annotation guidelines provided by Cripwell et al. (2023).

In Figure D, the screenshot of the human evaluation interface given to the annotators is shown.

E Scientific artifacts and licensing

In this work, we used the following scientific artifacts. BLOOM is licensed under The BigScience RAIL License. LLaMa2 is licensed under a commercial license [9]. GPT-3.5 is licensed under a commercial license [10]. Falcon is licensed under the FALCON 180B TII LICENSE VERSION 1.0 [11]. NLLB is licensed under CC-BY-NC-4.0 [12]. The

[9] https://ai.meta.com/resources/models-and-libraries/llama-downloads/

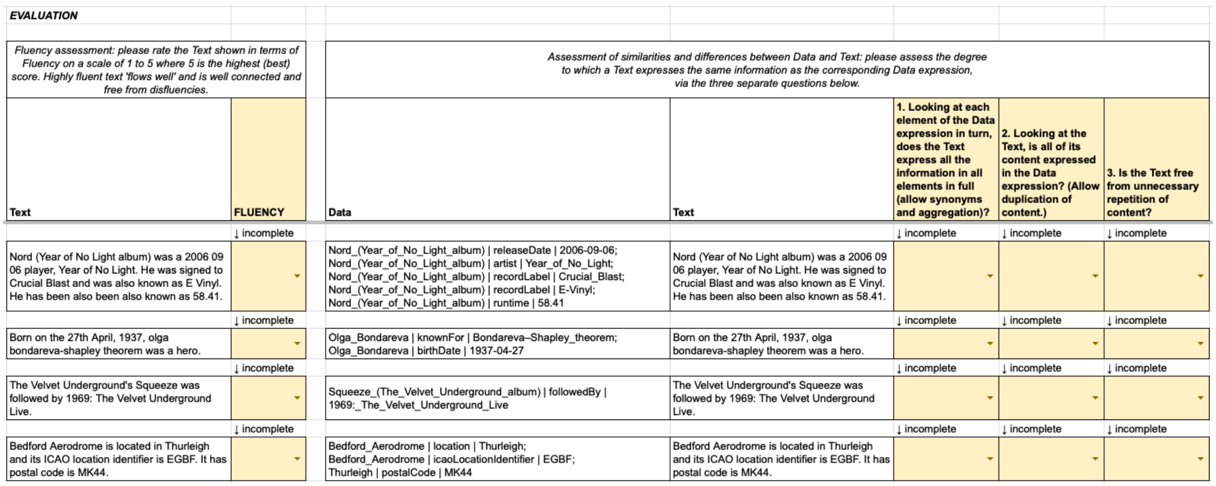
[10] https://openai.com/policies/terms-of-use

[11] https://huggingface.co/tiiuae/falcon-180B-chat/blob/main/ACCEPTABLE_USE_POLICY.txt

[12] https://huggingface.co/facebook/nllb-200-1.3B/blob/main/README.md

M	Prompt	Irish			Welsh			Breton		
		BERT-P↑	BERT-R↑	BERT-F1↑	BERT-P↑	BERT-R↑	BERT-F1↑	BERT-P↑	BERT-R↑	BERT-F1↑
GPT-3.5 (175B)	ZS MI	0.7574	0.7543	0.7555	0.7837	0.7796	0.7813	0.7768	0.7688	0.7722
	FS IC	0.7723	0.7661	0.7688	0.8057	0.7928	0.7989	***0.7979***	***0.7817***	***0.7892***
	ZS MI + GT	0.8115	0.8035	0.8071	0.8255	0.8253	0.8251	-	-	-
	FS IC + GT	***0.8149***	***0.8044***	***0.8093***	***0.8283***	***0.8259***	***0.8268***	-	-	-
	ZS MI + AT	0.8077	0.7973	0.8022	0.8217	0.8213	0.8212	0.7595	0.7384	0.7482
	FS IC + AT	0.8107	0.7984	0.8041	0.8253	0.8227	0.8237	0.7618	0.7379	0.749
	ZSMI+NLLB	0.7998	0.7824	0.7906	0.8149	0.7979	0.8057	-	-	-
	FSIC+NLLB	0.8025	0.7824	0.7919	0.8176	0.7977	0.807	-	-	-
BLOOM (176B)	ZS MI	0.6485	0.6282	0.6365	0.6166	0.6265	0.62	0.598	0.6181	0.6057
	FS IC	0.6857	0.6757	0.6797	0.7173	0.6928	0.7035	0.7178	0.699	0.7071
	ZS MI + GT	0.7432	0.7479	0.7442	0.7533	0.7641	0.7572	-	-	-
	FS IC + GT	*0.7829*	*0.7758*	*0.7786*	*0.7933*	*0.7921*	*0.7918*	-	-	-
	ZS MI + AT	0.7406	0.7435	0.7408	0.7514	0.7618	0.7552	0.7107	0.703	0.7054
	FS IC + AT	0.7758	0.7695	0.7718	0.7897	0.7893	0.7886	*0.7428*	*0.7247*	*0.7325*
	ZSMI+NLLB	0.6391	0.6241	0.6305	0.6497	0.6235	0.6353	-	-	-
	FSIC+NLLB	0.6525	0.6324	0.6416	0.6642	0.6308	0.6463	-	-	-
LLaMa2-chat (70B)	ZS MI	0.7051	0.6563	0.6781	0.7153	0.6742	0.6926	0.7214	0.6539	0.6843
	FS IC	0.7324	0.7278	0.7295	0.7272	0.7273	0.7265	0.7371	0.7101	0.7225
	ZS MI + GT	0.7909	0.79	0.7897	0.8025	0.8079	0.8043	-	-	-
	FS IC + GT	*0.8046*	*0.8007*	*0.8023*	*0.8168*	*0.8208*	*0.8184*	-	-	-
	ZS MI + AT	0.787	0.7847	0.7852	0.799	0.8054	0.8014	0.7461	0.7295	0.7368
	FS IC + AT	0.8	0.7949	0.797	0.8129	0.8176	0.8149	*0.7554*	*0.737*	*0.7453*
	ZSMI+NLLB	0.7834	0.7663	0.7739	0.7974	0.781	0.7881	-	-	-
	FSIC+NLLB	0.7949	0.7789	0.7862	0.8088	0.7931	0.8002	-	-	-
Falcon-chat (180B)	ZS MI	0.6961	0.6833	0.6885	0.7004	0.6854	0.6914	0.7232	0.6839	0.7013
	FS IC	0.7384	0.7397	0.7385	0.77	0.75	0.7589	0.7412	0.7119	0.7253
	ZS MI + GT	0.7656	0.7792	0.7712	0.7758	0.7967	0.7849	-	-	-
	FS IC + GT	*0.8029*	*0.8003*	*0.8012*	*0.8155*	*0.8221*	*0.8183*	-	-	-
	ZS MI + AT	0.7623	0.7743	0.7672	0.7743	0.7944	0.783	0.7307	0.726	0.7273
	FS IC + AT	0.7983	0.795	0.7962	0.8135	0.8197	0.8162	*0.7566*	*0.7387*	*0.7468*
	ZSMI+NLLB	0.7616	0.7594	0.7594	0.7748	0.774	0.7732	-	-	-
	FSIC+NLLB	0.7933	0.7784	0.7852	0.8094	0.7946	0.8012	-	-	-

Table 6: BERTScore results for **Irish**, **Welsh** and **Breton**. Maltese is not available in BERTScore. Highest score in each column for each language in bold, highest score for each model in italics. Number of parameters in brackets in column 1. ZS MI=Zero-Shot Minimal Instruction, FS IC=Few-Shot In Context, GT=Google Translate, AT=Alibaba Translate, NLLB=No Language Left Behind system.



EVALUATION

Fluency assessment: please rate the Text shown in terms of Fluency on a scale of 1 to 5 where 5 is the highest (best) score. Highly fluent text 'flows well' and is well connected and free from disfluencies.		*Assessment of similarities and differences between Data and Text: please assess the degree to which a Text expresses the same information as the corresponding Data expression, via the three separate questions below.*				
Text	FLUENCY	Data	Text	1. Looking at each element of the Data expression in turn, does the Text express all the information in all elements in full (allow synonyms and aggregation)?	2. Looking at the Text, is all of its content expressed in the Data expression? (Allow duplication of content.)	3. Is the Text free from unnecessary repetition of content?
	↓ incomplete			↓ incomplete	↓ incomplete	↓ incomplete
Nord (Year of No Light album) was a 2006 09 06 player, Year of No Light. He was signed to Crucial Blast and was also known as E Vinyl. He has been also been also known as 58.41.		Nord_(Year_of_No_Light_album) | releaseDate | 2006-09-06; Nord_(Year_of_No_Light_album) | artist | Year_of_No_Light; Nord_(Year_of_No_Light_album) | recordLabel | Crucial_Blast; Nord_(Year_of_No_Light_album) | recordLabel | E-Vinyl; Nord_(Year_of_No_Light_album) | runtime | 58.41	Nord (Year of No Light album) was a 2006 09 06 player, Year of No Light. He was signed to Crucial Blast and was also known as E Vinyl. He has been also been also known as 58.41.			
	↓ incomplete			↓ incomplete	↓ incomplete	↓ incomplete
Born on the 27th April, 1937, olga bondareva-shapley theorem was a hero.		Olga_Bondareva | knownFor | Bondareva–Shapley_theorem; Olga_Bondareva | birthDate | 1937-04-27	Born on the 27th April, 1937, olga bondareva-shapley theorem was a hero.			
	↓ incomplete			↓ incomplete	↓ incomplete	↓ incomplete
The Velvet Underground's Squeeze was followed by 1969: The Velvet Underground Live.		Squeeze_(The_Velvet_Underground_album) | followedBy | 1969:_The_Velvet_Underground_Live	The Velvet Underground's Squeeze was followed by 1969: The Velvet Underground Live.			
	↓ incomplete			↓ incomplete	↓ incomplete	↓ incomplete
Bedford Aerodrome is located in Thurleigh and its ICAO location identifier is EGBF. It has postal code is MK44.		Bedford_Aerodrome | location | Thurleigh; Bedford_Aerodrome | icaoLocationIdentifier | EGBF; Thurleigh | postalCode | MK44	Bedford Aerodrome is located in Thurleigh and its ICAO location identifier is EGBF. It has postal code is MK44.			

Figure 2: Screenshot of the human evaluation interface.

Model	Prompt	BERT		
		P↑	R↑	F1↑
GPT-3.5 (175B)	ZS MI	0.9555	0.9568	0.9555
	FS IC	***0.9588***	***0.9582***	***0.958***
BLOOM (176B)	ZS MI	0.9092	0.9234	0.9151
	FS IC	*0.938*	*0.937*	*0.9368*
LLaMa2-chat (70B)	ZS MI	0.9449	0.9465	0.9449
	FS IC	*0.9522*	*0.9535*	*0.9523*
Falcon-chat (180B)	ZS MI	0.9276	0.9379	0.9319
	FS IC	*0.9532*	*0.9543*	*0.9531*

Table 7: BERTScore results in **English**. Best score per column in bold, best score per model in italics. Number of model parameters in brackets. ZS MI=Zero-Shot Minimal Instruction, FS IC=Few-Shot In Context.

usage of the listed artifacts is consistent with their licenses.

M	Prompt	COMET ↑				
		Irish	Welsh	Maltese	Breton	English
GPT-3.5 (175B)	ZS MI	0.6606	0.7301	0.6378	0.6772	0.8261
	FS IC	0.6994	0.7521	0.6425	***0.6962***	***0.8306***
	ZS MI + GT	0.7387	0.7918	0.676	-	-
	FS IC + GT	*0.7431*	***0.7939***	***0.6739***	-	-
	ZS MI + AT	0.7205	0.7776	0.6584	0.5698	-
	FS IC + AT	0.7279	0.7796	0.6557	0.5711	-
	ZSMI+NLLB	0.7155	0.7513	0.6583	-	-
	FSIC+NLLB	0.715	0.7542	0.6584	-	-
BLOOM (176B)	ZS MI	0.4525	0.4152	0.4426	0.428	0.7186
	FS IC	0.4523	0.4837	0.5401	0.5242	*0.7799*
	ZS MI + GT	0.6569	0.7015	0.6125	-	-
	FS IC + GT	*0.7063*	*0.7512*	*0.6426*	-	-
	ZS MI + AT	0.6459	0.6865	0.602	0.522	-
	FS IC + AT	0.6884	0.7386	0.6274	*0.5596*	-
	ZSMI+NLLB	0.6327	0.6686	0.6027	-	-
	FSIC+NLLB	0.6812	0.7191	0.6289	-	-
LLaMa2-chat (70B)	ZS MI	0.4761	0.4775	0.5403	0.416	0.7962
	FS IC	0.5959	0.541	0.5923	0.4866	*0.8211*
	ZS MI + GT	0.7204	0.7662	0.655	-	-
	FS IC + GT	*0.7381*	*0.7856*	*0.6689*	-	-
	ZS MI + AT	0.7005	0.7546	0.6386	0.5583	-
	FS IC + AT	0.7185	0.7722	0.6492	*0.5696*	-
	ZSMI+NLLB	0.688	0.7303	0.6369	-	-
	FSIC+NLLB	0.7092	0.7495	0.648	-	-
Falcon-chat (180B)	ZS MI	0.5393	0.5437	0.5331	0.4854	0.765
	FS IC	0.6182	0.6566	0.599	0.5599	*0.8229*
	ZS MI + GT	0.7063	0.7487	0.6363	-	-
	FS IC + GT	***0.7457***	*0.7922*	*0.6709*	-	-
	ZS MI + AT	0.6866	0.7371	0.6227	0.5519	-
	FS IC + AT	0.7257	0.7817	0.6534	*0.5731*	-
	ZSMI+NLLB	0.6809	0.7186	0.622	-	-
	FSIC+NLLB	0.7154	0.7546	0.6498	-	-

Table 8: COMET results for **Irish**, **Welsh**, **Maltese**, **Breton**, and **English**. COMET scores are between 0 and 1. Highest score in each column for each language in bold, highest score for each model in italics. Number of parameters in brackets in column 1. ZS MI=Zero-Shot Minimal Instruction, FS IC=Few-Shot In Context, GT=Google Translate, AT=Alibaba Translate, NLLB=No Language Left Behind system.

Zero-Shot Minimal Instruction	
Template:	Write the following triples as fluent English | Irish | Welsh | Maltese | Breton text. Triples: """ {set of triples in the format *subject predicate object* and each triple in a new line} """ Text: [MODEL]

Table 9: Template of the Zero-Shot Minimal Instruction prompt.

Few-Shot In Context	
Template:	Write the following triples as fluent English | Irish | Welsh | Maltese | Breton text. Triple 1: """ {set of triples in the format *subject predicate object* and each triple in a new line} """ Text 1: {verbalisation of Triple 1} ## Triple 2: """ {set of triples in the format *subject predicate object* and each triple in a new line} """ Text 2: {verbalisation of Triple 2} ## Triple 3: """ {set of triples in the format *subject predicate object* and each triple in a new line} """ Text 3: [MODEL]
English, Irish, and Breton Triples:	Triple 1: Adolfo_Suárez_Madrid–Barajas_Airport runwayName "14R/32L" Triple 2: American_Journal_of_Mathematics abbreviation "Am. J. Math." American_Journal_of_Mathematics firstPublicationYear 1878 American_Journal_of_Mathematics issnNumber "1080-6377"
English texts:	Text 1: 14R/32L is the runway name of Adolfo Suárez Madrid-Barajas Airport. Text 2: The American Journal of Mathematics was first published in 1878 and is also known by the abbreviated title of Am. J. Math. It has an ISSN number of 1080-6377.
Irish texts:	Text 1: 14R/32L is ainm do rúidbhealach Aerfort Adolfo Suárez Madrid-Barajas Text 2: Foilsíodh an American Journal of Mathematics don chéad uair in 1878 agus aithnítear leis an ainm giorraithe Am. J. Math. chomh maith é. Tá an uimhir ISSN 1080-6377 aige.
Breton texts:	Text 1: Anv leurenn bradañ aerborzh Adolfo Suárez Madrid-Barajas zo 14L/32R. Text 2: Finland zo bro ar Finniz hag hini ar skorndorrer Aleksey Chirikov bet savet e chanter-bigi Arctech en Helsinki.
Maltese and Welsh Triples:	Triple 1: Albennie_Jones birthPlace Errata,_Mississippi Triple 2: GMA_New_Media industry Entertainment GMA_New_Media type Media_company GMA_New_Media product World_Wide_Web
Maltese texts:	Text 1: Albennie Jones twieldet f'Errata Mississippi. Text 2: GMA New Media hija kumpanija tal-midja tal-industrija tad-divertiment li toffri servizzi li jikkonċernaw il-World Wide Web.
Welsh texts:	Text 1: Ganed Albennie Jones yn Errata, Mississippi. Text 2: Mae GMA New Media yn gwmni cyfryngau yn y diwydiant adloniant sy'n cynnig gwasanaethau sy'n ymwneud â'r We Fyd Eang.

Table 10: Few-Shot In Context prompt. **Top** Template of the prompt. **Center** Examples' triple set and texts in English, Irish, and Breton. **Bottom** Examples' triple set and texts in Maltese and Welsh.