TARJAMAT: Evaluation of Bard and ChatGPT on Machine Translation of Ten Arabic Varieties

Karima Kadaoui[λ,⋆] **Samar M. Magdy**[λ,⋆] **Abdul Waheed**[λ,⋆] **Md Tawkat Islam Khondaker**[ξ,⋆]
Ahmed Oumar El-Shangiti[λ] **El Moatez Billah Nagoudi**[ξ] **Muhammad Abdul-Mageed**[ξ,λ,⋆]
[ξ] Deep Learning & Natural Language Processing Group, The University of British Columbia
[λ]Department of Natural Language Processing & Department of Machine Learning, MBZUAI
{karima.kadaoui,samar.magdy,abdul.waheed,ahmed.oumar}@mbzuai.ac.ae
{tawkat@cs,moatez.nagoudi,muhammad.mageed}@ubc.ca

Abstract

Despite the purported multilingual proficiency of instruction-finetuned large language models (LLMs) such as ChatGPT and Bard, the linguistic inclusivity of these models remains insufficiently explored. Considering this constraint, we present a thorough assessment of Bard and ChatGPT (encompassing both GPT-3.5 and GPT-4) regarding their machine translation proficiencies across ten varieties of Arabic. Our evaluation covers diverse Arabic varieties such as Classical Arabic (CA), Modern Standard Arabic (MSA), and several country-level dialectal variants. Our analysis indicates that LLMs may encounter challenges with dialects for which minimal public datasets exist, but on average are better translators of dialects than existing commercial systems. On CA and MSA, instruction-tuned LLMs, however, trail behind commercial systems such as Google Translate. Finally, we undertake a human-centric study to scrutinize the efficacy of the relatively recent model, Bard, in following human instructions during translation tasks. Our analysis reveals a circumscribed capability of Bard in aligning with human instructions in translation contexts. Collectively, our findings underscore that prevailing LLMs remain far from inclusive, with only limited ability to cater for the linguistic and cultural intricacies of diverse communities.

1 Introduction

Large language models (LLMs) finetuned to follow instructions (Wei et al., 2021; Wang et al., 2022; Ouyang et al., 2022) have recently emerged as powerful systems for handling a wide range of NLP tasks. In accordance with the scaling law (i.e., pretraining larger models will continue to result in better performance) (Kaplan et al., 2020), a number of LLMs such as GPT-3 (Brown et al., 2020), Chinchilla (Hoffmann et al., 2022), Claude (Anthropic, 2023), ChatGPT[1] (OpenAI, 2022), GPT-4 (OpenAI, 2023), and Bard (Google, 2023) have been introduced. Most of these models, however, are ‘closed’. That is, little-to-no information about them is known. This includes details about model architectures, pretraining data, languages involved, and training configurations. LLMs are also expensive both to pretrain and deploy. To alleviate these concerns, ‘open’ LLMs such as BLOOM (Scao et al., 2022), LLaMA-1 (Touvron et al., 2023a), Falcon (Almazrouei et al., 2023), and LLaMA-2 (Touvron et al., 2023b) were introduced. These more open models can facilitate research and (non-)commercial deployment.

Figure 1: Experimental setup for our evaluation. We evaluate multiple language models on different Arabic varieties.

In spite of drawbacks such as their closed nature, computational costs (Dasgupta et al., 2023), and biases they exhibit (Ferrara, 2023), closed LLMs remain attractive primarily due to their remarkable performance (Bang et al., 2023a; Laskar et al., 2023a). It is thus important to fully understand the full capabilities of these closed models. Although there has been a recent flurry of works attempting to evaluate ability of LLMs to carry out NLP tasks, many of these models remain opaque. This is especially the case when it comes to understanding how LLMs fare on different varieties and dialects of several popular languages and on vital tasks such as machine translation (MT). For example, the extent to which LLMs can handle MT from Arabic varieties into other languages is unknown.

⋆Equal contribution

[1]In this work, we refer `gpt-3.5-turbo` as ChatGPT.

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 52–75
December 7, 2023

Another challenge is how more recent models such as Google's Bard are yet to be evaluated and understood. Bard was released in 41 different languages, which makes it a particularly attractive target for MT evaluation. This is also the case given Google's strong history of investment in MT (Wu et al., 2016a). In this work, we offer a thorough evaluation of LLMs on MT from major Arabic varieties into English (Figure 1). Namely, we evaluate ChatGPT, GPT-4, and Bard on MT of ten Arabic varieties into English. Since there are usually concerns about downstream evaluation data leaking into LLM pretraining, which involves data collected from the web, we benchmark the models on new test sets that we manually prepare for this work. Our evaluation targets diverse varieties of Arabic. Namely, we evaluate on Classical Arabic (CA), Modern Standard Arabic (MSA), and several country-level Arabic dialects such as Algerian and Egyptian Arabic (Section 3).

Bard provides three different drafts for each text input we ask it to translate. Contents of the three drafts are diverse, providing us with excellent contexts to analyze the degree to which the model adheres to our prompts. We leverage these contexts to carry out a human evaluation study investigating the *helpfulness* of the model, allowing us to reveal a number of Bard's limitations. We carefully analyze these limitations against the different Arabic varieties we target, thus affording even better understanding of the model's ability to translate from Arabic.

Overall, our work offers the following contributions:

(i) We offer a detailed MT evaluation of instruction finetuned LLMs on ten diverse varieties of Arabic.

(ii) To the best of our knowledge, our work is the first to assess performance of Bard on NLP tasks in any language, and on Arabic MT in particular.

(iii) We introduce a new manually created multi-Arabic dataset for MT evaluation that has never been exposed to any existing LLM.

(iv) We extensively evaluate Bard through a human study to analyze its behavior in terms of *helpfulness*. We examine how well the model follows human instructions when tasked with translating across ten different Arabic varieties.

The rest of the paper is organized as follows: In Section 2, we review previous research evaluating LLMs on NLP tasks in general and MT in particular. In Section 3, we introduce our newly developed multi-Arabic MT dataset. In Section 4, we describe our evaluation methods. In Section 5, we present our results and the main findings obtained from comparing ChatGPT and Bard to various commercial MT products. In Section 6, we present our human study analyzing Bard's helpfulness, particularly in terms of its ability to follow human instructions in MT. We conclude in Section 7.

2 Related Work

Evaluation of ChatGPT and Other LLMs. A growing body of literature has focused on evaluating ChatGPT and other LLMs on NLP tasks. Laskar et al. (2023a) find ChatGPT effective on many tasks. Other works find it either on par with supervised models (Ziems et al., 2023) or in some cases (e.g., sequence tagging) falling behind these models (Qin et al., 2023). Both Jiao et al. (2023) and Ogundare and Araya (2023) find that GPT-4 is competitive with commercial systems for high-resource languages but lags behind for low-resource languages. Bang et al. (2023b) find a similar pattern for ChatGPT. Guerreiro et al. (2023) find complex translation scenarios, such as in the low-resource setting, to be prone to hallucination. Peng et al. (2023) demonstrate that ChatGPT can surpass Google Translate on many translation pairs, but Zhu et al. (2023) show it is outperformed by NLLB (NLLB et al., 2022) on at least 83% of the English-centric pairs they study. Wang et al. (2023); Karpinska and Iyyer (2023), however, show that ChatGPT can match the performance of fully supervised models for document-level translation.

Peng et al. (2023) find that adding task and domain-specific information in the prompt can improve the robustness of the MT system, which corroborates the findings by Gao et al. (2023). Huang et al. (2023) propose a prompting technique called cross-lingual-thought prompting (XLT) to improve cross-lingual performance for a wide range of tasks, including MT. Similarly, Lu et al. (2023b) asks ChatGPT to correct its own mistakes as a way to improve the model's translation quality. Lu et al. (2023a) propose Chain-of-Dictionary (CoD) prompting to solve rare word translation issues. Prompting with CoD improves the performance of ChatGPT for both X-En and En-X language directions.

Evaluation of ChatGPT on Arabic. Khondaker et al. (2023) evaluate ChatGPT and other contemporary LLMs such as BloomZ (Muennighoff et al., 2022) in few-shot settings (0, 1, 3, 5, and 10) on four X-Arabic and two code-mixed Arabic-X language sets. They show that providing in-context examples to ChatGPT achieves comparable results to a supervised baseline. Alyafeai et al. (2023) evaluate ChatGPT and GPT-4 on $4,000$ Arabic-English sentence pairs from Ziemski et al. (2016) and find a supervised SoTA model to outperform ChatGPT and GPT-4 by a significant margin. These works, however, only consider a limited number of Arabic varieties. They also do not conduct a thorough analysis of the LLMs for MT. Additionally, none of the works evaluate Bard. Our work bridges these gaps by performing a comprehensive evaluation of these systems on a wide range of Arabic varieties. We also conduct our study on novel in-house data that we guarantee no leakage for (i.e., our data cannot have been seen by ChatGPT, GPT-4, or Bard since we create the data for this work). Other works have focused on evaluating smaller-sized Arabic language models (Abu Farha and Magdy, 2021; Inoue et al., 2021; Alammary, 2022), including on recent benchmarks (Nagoudi et al., 2023; Elmadany et al., 2023).

Arabic MT. There are several works on Arabic MT itself, including rule-based (Bakr et al., 2008; Mohamed et al., 2012; Salloum and Habash, 2013), statistical (Habash and Hu, 2009; Salloum and Habash, 2011; Ghoneim and Diab, 2013), and neural (Junczys-Dowmunt et al., 2016; Almahairi et al., 2016; Durrani et al., 2017; Alrajeh, 2018). While these systems focus on MSA, others target Arabic dialects (Zbib et al., 2012; Sajjad et al., 2013; Salloum et al., 2014; Guellil et al., 2017; Baniata et al., 2018; Sajjad et al., 2020; Farhan et al., 2020; Nagoudi et al., 2021, 2022a). We provide a more detailed review of related literature in Appendix A, with a summary in Table 7.

3 Coverage and Datasets

3.1 Arabic Varieties

Our goal is to provide a comprehensive evaluation of MT on ChatGPT, GPT-4, and Google Bard, focusing on their performance across ten different varieties of Arabic. These can vary across *time* (i.e., old vs. modern day) and *space* (e.g., country-level geography) as well as their *sociopragmatic* functions (e.g., standard use in government communication vs. everyday street language). Before introducing our dataset, we provide a brief background about Arabic and its varieties. Arabic, the collection of languages spoken by approximately 450 million people across the Arab world, encompasses a broad spectrum of varieties. Classical Arabic (CA) is known as Quranic Arabic, the language of the Quran (Rabin, 1955), and has emerged from the medieval dialects of the Arab tribes. It was spoken early in Mecca around $1,500$ years ago in the sixth or seventh century AD. CA is considered the most eloquent form of Arabic and is preserved notably in the Holy Quran and pre-Islamic epic poems (Versteegh, 2014). It is often described as exhibiting archaic words, figurative speech, and rhyming sentences that are no longer (or less frequently) used in MSA and dialectal Arabic varieties. Modern Standard Arabic (MSA) (Holes, 2004), on the contrary, is deeply rooted in CA that has been lightened to a great extent to encompass the modern uses in Modern literature, poetry and official statements. MSA additionally serves as the standardized language for formal events, news broadcasts, sermons, and formal communication. We now explain how we acquire our dataset for each Arabic variety.

Variety	Example with English Translation
EGY	ماحنا لو فضلنا مدارين و مستخبين هنموت من الخوف. And if we keep hiding, we're going to die out of fear
JOR	أنا مش مستخف فيه و لا يمكن استخف فيه مهما كان بضل ابوي I do not and cannot underestimate him; he is still my father, no matter what.
MAU	شوف آن شي كامل ادخلتو ماتيت انركيلي فيه خالك اللا القدام. Look, whenever I'm in, I never take a step back; I only go forward.
YEM	ركزت لي نفطة تفتيش، في الباب، بتفتش ذي داخلي و ذي خارجي. I set up a checkpoint at the door to screen anyone who comes in or out.

Table 1: Example sentences from some of the Arabic varieties in our new translation evaluation dataset. See Appendix Table 16 for remaining varieties.

Prompt	Template	BLEU
ENG	Translate the following Modern Standard Arabic (MSA) sentence into English	48.48
MSA	ترجم الجملة العربية الفصحى العصرية التالية الي اللغة الإنجليزية	47.92
ENG (elaborate)	I want you to act as an expert translator. You will translate Modern Standard Arabic (MSA) sentences into English. I will give you a Modern Standard Arabic (MSA) input, and you will translate it into English and keep the same semantic meaning. Please translate this Modern Standard Arabic (MSA) text into English	46.17

Table 2: Performance of ChatGPT on the MSA→English translation task. Our concise English prompt outperforms other prompts in BLEU score.

3.2 Datasets

CA. We manually curate 200 sentences from the Open Islamic Texts Initiative (OpenITI) (Nigst et al., 2020) dataset, namely from the latest 2022.16 version. It includes a collection of premodern Arabic works featuring a comprehensive library of 10,342 books. The sentences were chosen based on a set of specified criteria: Initially, we identify books originating from the first and second-century Anno Hegirae (in the year of the Hijra), excluding those written after this period. Then we compile a collection of 15 distinctive books, including notable works like Abdullah Ibn AlMuqfaa's "Al-Adab Al-Kabir" and "Al-Adab Al-Saghir", Mohamed Idis Al-Shafi's "Al-Umm", "Al-Risala", and "Al-Adab Wal-Muraa", among others. We subsequently extract sentences of a minimum of ten words. We provide the list of the 15 books we sample from in Appendix B (Table 9).

MSA. We collect a total of 200 sentences from current event news picked from two online news websites: Aljazeera[2] and BBC Arabic[3]. The curated sentences showcase various news genres, including political, social, and sports.

Various Dialects. We manually select a dataset of dialectal Arabic from an in-house project where we transcribe TV series collected from YouTube videos belonging to Arabic dialects. Again, we use 200 sentences from each dialect, resulting in a total of 1,600 sentences across eight dialects, each transcribed and translated by their respective native speakers. The dialects belong to North African countries such as Algeria, Morocco, and Mauritania; Gulf area dialects, namely Emirati; Levantine Arabic (focusing on Palestinian and Jordanian); and Egyptian Arabic.

For all varieties, we collect sentences that are *at least ten words* long. We present one sample from some of the dataset in Table 1. Statistics of the datasets across the Arabic varieties is presented in Appendix B (Table 8).

4 Methodology

4.1 Prompt Design

The term *prompt* refers to the set of instructions used to program an LLM with a goal to steer and enhance its purpose and capabilities (White et al., 2023). Prompts can influence subsequent interactions with the model as well as its generated outputs. Therefore, it is important to clearly identify the right prompts to obtain the desired outcome for a particular task. To determine the right prompt for our translation task, we set up a pilot experiment that we now describe.

Pilot experiment. In our pilot experiment, we investigate three prompt candidates. To limit the search space, we perform this experiment only with ChatGPT. We experiment with both Arabic and English prompts to *concisely* instruct ChatGPT to translate from an Arabic variety into English, again restricting our search space to MSA as a variety that is known to overlap with other varieties at all linguistic levels (Abdul-Mageed et al., 2020; Habash, 2022). We also experiment with an *elaborate* English prompt that clearly defines the role and the objective of ChatGPT before asking the model to carry out the translation task. We then evaluate the performance of ChatGPT on 100 MSA$\rightarrow$English samples. We present the prompt templates and the corresponding performance we acquire in Table 2.

Evaluation. As evident, the concise English prompt outperforms the other two prompts, including the Arabic counterpart (by 1$\sim$2 BLEU scores). This result substantiates findings in prior works (Khondaker et al., 2023; Lai et al., 2023) regarding the superiority of English prompts on ChatGPT over non-English prompts. Therefore, in the rest of the paper we employ the concise and direct English prompt to conduct our experiments.

4.2 *N*-Shot Experiments

We run ChatGPT MT generation under 0-shot, 1-shot, 3-shot, and 5-shot settings. For a particular translation task, we always select the samples for these in-context learning experiments from the same set of training examples. This means that for a k-shot setting, we make sure that if a training sample is selected then it will also be selected for n-shot settings where $n > k$. We generate translation with ChatGPT (`gpt-3.5-turbo`[4], an optimized version of GPT-3.5 series) by setting the temperature to 0.0 to ensure *deterministic and reproducible results*. In addition, we restrict the maximum token length to 512 for all the generation tasks. For GPT-4, we use the web interface for MT generation under 0-shot and 5-shot settings. For Bard[5], we use the web interface but opt out of gen-

[2] https://aljazeera.net/news
[3] https://bbc.com/arabic
[4] Snapshot of `gpt-3.5-turbo` from June 13th 2023.
[5] Update from - 2023.07.13

erating any few-shot response because it lacks an API and its outputs can be problematic requiring intensive manual preprocessing (Section 6).

4.3 Evaluation and Baselines

Evaluation metrics. Different evaluation metrics are usually employed to automatically evaluate MT systems. These metrics are often based on word overlap and/or context similarity between references and model outputs. In our work, we employ both types of metrics to evaluate the quality of various translation systems that we consider in our study. Namely, we use BLEU (Papineni et al., 2002), COMET (Rei et al., 2020a), ChrF (Popović, 2015), ChrF++, and TER (Snover et al., 2006). We provide a detailed description of each metric in Appendix 4.1.

Baselines. We compare instruction-tuned LLMs to a number of MT systems, including both commercial services (Amazon, Google, and Microsoft) as well as the supervised NLLB-200 system (NLLB et al., 2022)[6]. We provide more details about each of these systems in Appendix 4.2.

5 Results and Discussion

We evaluate all models on X-English translation direction where X is an Arabic variety (MSA and CA). As mentioned earlier, we evaluate LLMs (ChatGPT, GPT-4, and Bard) in n-shot settings. We report BLEU, COMET, and ChrF++ in Table 3. We report additional metrics in Appendix C. We summarize our main findings here.

Is GPT-4 better than ChatGPT? *In most cases, yes*. GPT-4 consistently outperforms ChatGPT on many dialects and varieties. However, for JOR and UAE, ChatGPT 0-shot performs better than 0-shot GPT-4. Overall, on average, GPT-4 0-shot outperforms ChatGPT 0-shot by $1 \sim 3$ points on all metrics. Additionally, GPT4 in 0-shot setting is on par with ChatGPT in the 5-shot setting. When comparing ChatGPT with GPT-4 under 5-shot setting, we observe that ChatGPT substantially closes the performance gap, even outperforming GPT-4 in 6 out of 10 varieties in terms of BLEU score. Although GPT-4 marginally outperforms ChatGPT on average BLEU score, *this result shows that by providing few-shot examples, it is possible for ChatGPT to achieve comparable performance to GPT-4 on Arabic MT.*

[6]For NLLB-200, we use the distilled 1.3B

Is ChatGPT/GPT4 better than Bard? *In most cases, yes*. For fairness, we compare Bard, ChatGPT, and GPT-4 only under the 0-shot condition. In the majority of the varieties, either ChatGPT or GPT-4 outperforms the best Bard draft (i.e., Draft 1). Our results show that Bard is better than both of these models in only three cases (i.e., CA, EGY and JOR). Overall, GPT-4 ranks best (BLEU score at 23.12), followed by ChatGPT (21.77 BLEU points), which in turn is followed by Bard (20.47 BLEU points).

Is ChatGPT/GPT4 better than commercial systems? *Yes, but only on dialects*. We evaluate three commercial translation systems, namely, Amazon, Microsoft, and Google Translate. Among commercial systems, we find Google Translate to outperform other commercial systems across all varieties except YEM. The average score for Google Translate is 22.29/64.89/43.11 (BLEU/COMET/ChrF++) compared to 18.80/63.68/41.55 and 17.77/62.85/39.76 for Microsoft and Amazon systems, respectively.

From our evaluation results in Table 3, we observe that commercial systems are better at translating CA and MSA but fail to produce high-quality translations when it comes to dialectal Arabic. ChatGPT and GPT-4 in 0-shot and few-shot settings are on par or better than the best-performing commercial system (i.e., Google Translate) for all Arabic dialects except JOR. The average BLEU score of ChatGPT and GPT-4 in few-shot setting is 23.62 (5-shot) and 13.64 (5-shot), respectively, compared to 2.29 for Google Translate. However, we notice that Google Translate outperforms ChatGPT and GPT-4 on MSA by a significant margin (while it stays behind on other dialects). Hence, we conclude that *ChatGPT and GPT-4 are better translators of Arabic dialects than the commercial Google Translate system*. We find similar patterns in other metrics.

Is ChatGPT/GPT-4 better than the supervised baseline? *Yes, it is*. We evaluate NLLB (NLLB et al., 2022) as the supervised baseline, finding both ChatGPT and GPT-4 able to outperform this baseline in the 0-shot setting. The average BLEU score for NLLB is 12.97 compared to 21.77 and 23.12 of ChatGPT and GPT-4 under 0-shot settings, respectively. Similar to the commercial systems, the supervised baseline (NLLB) does well on MSA and is on par with ChatGPT and GPT-4. However, both ChatGPT and GPT-4 outperform it

Met	Var/M	ChatGPT				GPT-4		Bard				NLLB	NLLB	Amazon	MST	GT
		0-shot	1-shot	3-shot	5-shot	0-shot	5-shot	D1	D2	D2	Avg	(SB)	(Dia)			
BLEU	CA	11.27	12.02	12.22	12.52	11.79	11.36	12.32	10.43	12.39	11.71	7.32	-	11.35	11.96	**14.30**
	MSA	42.85	44.11	44.29	44.81	43.18	43.66	37.23	33.23	36.18	35.55	41.34	-	46.76	47.36	**66.01**
	ALG	14.48	16.41	17.16	17.31	18.37	**17.83**	15.24	11.67	12.58	13.16	7.27	-	10.08	11.67	11.93
	EGY	19.96	21.00	21.38	**21.74**	21.15	21.49	21.33	19.39	20.91	20.54	11.12	13.87	14.95	16.64	18.09
	JOR	25.74	26.75	27.63	26.82	24.57	25.26	26.93	23.48	25.09	25.17	13.07	18.5	21.56	21.71	**29.35**
	MAU	8.52	8.96	9.27	9.05	9.19	**9.87**	6.11	4.25	2.37	4.24	3.48	-	7.21	6.89	7.67
	MOR	27.15	28.19	28.86	29.80	32.90	**33.32**	31.59	30.84	31.25	31.23	10.45	19.47	12.76	14.25	16.94
	PAL	29.47	29.37	31.62	31.56	**31.97**	30.48	22.57	20.59	24.25	22.47	14.98	12.56	21.75	24.23	25.78
	UAE	24.20	24.61	24.55	26.17	23.86	**26.91**	21.93	19.61	21.29	20.94	11.27	-	16.85	19.05	19.56
	YEM	14.03	15.13	16.24	**16.44**	14.27	16.22	9.46	6.38	5.33	7.06	9.41	12.56	14.41	14.23	13.25
	Avg	21.77	22.66	23.32	23.62	23.12	**23.64**	20.47	17.99	19.16	19.21	12.97	15.39	17.77	18.80	22.29
COMET	CA	70.11	70.08	70.01	70.24	**71.47**	70.95	68.29	67.04	68.65	67.99	58.87	-	63.03	63.16	66.37
	MSA	85.87	86.14	86.22	86.24	86.32	86.22	80.21	80.00	80.44	80.22	84.76	-	86.15	85.70	**87.23**
	ALG	62.69	63.77	63.98	63.85	65.06	**65.52**	60.90	55.62	59.72	58.75	49.88	-	54.55	56.48	55.33
	EGY	72.41	73.15	74.20	73.96	74.14	**74.91**	71.50	68.20	71.30	70.33	61.15	63.81	64.24	65.59	68.41
	JOR	74.46	75.20	75.52	75.27	76.37	**76.50**	74.19	70.65	72.65	72.50	60.25	65.05	67.33	70.46	71.83
	MAU	58.37	58.99	60.35	60.66	59.24	**62.13**	52.53	46.38	50.41	49.77	48.50	-	52.37	51.45	51.58
	MOR	69.36	69.64	70.58	70.73	73.94	**73.95**	72.12	70.60	71.82	71.51	53.23	62.74	54.50	51.89	56.55
	PAL	74.59	74.94	75.40	75.51	**76.62**	76.19	69.37	67.78	69.94	69.03	60.57	59.04	65.80	68.54	68.69
	UAE	69.64	69.62	69.80	70.80	**72.93**	72.38	66.71	63.08	66.12	65.30	54.57	-	59.40	61.74	61.57
	YEM	64.48	65.41	66.09	65.88	62.47	**68.77**	58.34	55.35	56.89	56.86	57.01	59.04	61.09	61.75	61.32
	Avg	70.20	70.69	71.22	71.31	71.86	**72.75**	67.42	64.47	66.79	66.23	58.88	61.94	62.85	63.68	64.89

Table 3: Results in BLEU, and COMET scores. Higher is better unless otherwise specified by ↓. Average represents the mean across all varieties. Three drafts (D1, D2, D3) from Bard are reported individually and averaged. NLLB is our MSA-based supervised baseline; NLLB (Dia) is dialect-specific. Abbreviations: SB - supervised baseline, Dia - dialect, Var - varieties, M - model, MST - Microsoft Translation, GT - Google Translate. Best results are in **bold**.

on dialectal translation by a significant margin.

Is NLLB with dialects as source better than vanilla NLLB? *Yes, it mostly is when the dialects match*. Our supervised baseline, NLLB, takes the dialects of the source into consideration. For example, both JOR and PAL dialects in NLLB can be defined as South Levantine, i.e., *(JOR, PAL)→South Levantine*. In addition, source dialects like EGY and MOR can be defined in their actual forms, while YEM can be defined as Taizzi. The column *NLLB (Dia)* in Table 3 provides BLEU score where the NLLB model treats the input as a particular dialect. We find that when the actual dialect matches the appropriate mapping with this NLLB source dialect, we acquire performance. One exception is the case of PAL, where NLLB does poorly compared to MSA.

Is Bard a good instruction following model? *Not always.* We evaluate Bard for our translation using the web interface[7]. We find that Bard can fail to follow the instructions we prompt it with. We further discuss and describe this in Section 6. Bard often provides the main translation output within double quotes (""), which we extract semi-automatically.[8] Additionally, Bard provides three different drafts. We report results for each draft independently, as well as the average of all three drafts in our results.

Are instruction following models better at dialect translation? *In most cases? Yes.* In order to clearly see performance on dialects, we exclude CA and MSA results and report the average performance of the models on the various dialects as reported in Table 4. We observe that GPT-4 at its 5-shot setting is the best model on dialects. Although commercial systems fare well on CA and MSA, their performance degrades on dialects. For example, the gap between the best performing commercial system (Google Translate) and the best instruction-tuned model (GPT-4 5-shot) across the various dialects races to 4.85 from 1.35 in terms of average BLEU score.

Do diacritics affect translation? *Yes, in most cases they do.* Although in most real-world use, native speakers do not usually employ diacritics,

[7] https://bard.google.com/

[8] In order to keep sufficient information to study model behavior, we collect and save all output from Bard (including explanations of translations). Even when we try to prompt Bard to restrict its output to target translation, it did not follow our instructions.

Metric	CGPT 0-shot	CGPT 5-shot	GPT-4 0-shot	GPT-4 5-shot	Bard	NLLB	GT
BLEU	20.44	22.36	22.03	**22.67**	19.40	10.13	17.82
COMET	68.25	69.58	70.10	**71.29**	65.71	55.65	61.91
ChrF++	43.71	44.70	44.98	**45.44**	36.23	28.64	39.33
TER↓	77.08	72.23	74.07	**71.51**	83.62	101.38	79.38

Table 4: Average scores across eight dialects, excluding MSA and CA. Higher is better unless specified by ↓. Best results are in **bold**.

some Arabic texts (especially those written in CA) do make use of diacritic markers. We were inquisitive about the effect of diacritics on the translation task across the different systems and so carry out a limited study of any such effect. To this end, we collect and manually translate 50 new CA sentences that are fully diacritized. The sentences conform to the identical selection criteria as those utilized within the study, specifically with regard to their length and as they originate from the first and second centuries AH books. We make a copy of this set and remove diacritics, and then independently feed both the diacritized and undiacritized versions to all the systems that we evaluate in this work. As shown in Table 5, we find most systems to work better when we remove diacritics. However, we also observe that some systems provide the same output regardless of whether the input is diacritized or not. This prompts us to conduct a quick analysis on a list of 20 word pairs of heterophonic homographs, i.e., words with the same spelling that change meaning and pronunciation according to the diacritics. We provide this list in Appendix 12 (Table 14). An example of such a pair is كَتَبَ – *he wrote* and كُتُبٌ – *books*. For this analysis, we perform single word translation by all the systems to ensure that the intended meaning cannot be retrieved from context, but rather solely based on changes in the diacritics. We find that Google Translate and Microsoft Translation provide the same meaning for both words of each pair, while the rest of the systems show different outputs when diacritics change.

Robustness. We also run a series of bootstrapping experiments that confirm the robustness of the results we acquire from the different models. We describe these experiments in Appendix 3.2.

Met	Mo/Var	CGPT	GPT-4	Bard D1	Bard Avg	NLLB	Amazon	MST	GT
BLEU	CA	**23.57**	23.81	22.94	23.01	**16.13**	17.50	**20.13**	26.61
	CA*	23.46	**24.45**	**25.39**	**24.25**	13.61	**18.66**	**20.13**	**24.92**
COMET	CA	74.38	75.07	73.23	73.27	**64.06**	63.98	65.39	72.04
	CA*	**75.75**	**76.71**	**76.01**	**75.56**	61.82	**66.01**	**66.60**	**73.76**

Table 5: Effect of diacritics on translation. CA* is without diacritics. Other metrics and bootstrapped results are reported in Appendix 3.3 (Tables 12 and 13).

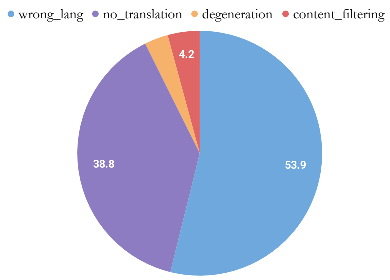
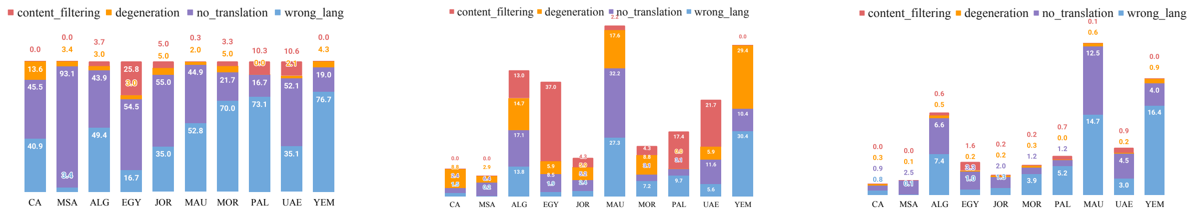


Figure 2: Distribution of Bard helpfulness errors when it fails to follow our prompts.

6 Human Analysis of Bard Helpfulness

Our experience working with Bard reveals that the model does not always follow human instructions. For this reason, we decided to carry out a human study to assess Bard's helpfulness. We define *helpfulness* here simply as the model's ability to follow human instructions. For each variety of Arabic, we task two native speakers of Arabic with familiarity with the dialects to assign one tags from the set {`wrong_lang`, `no_translation`, `degeneration`, `content_filtering`} to the model responses. We develop this tagset based on a bottom-up approach where we let the categories emerge from the data. Although this tagset may not be exhaustive, we find it to reasonably capture errors we identify with model responsiveness to instructions. Each of the two annotators manually label each draft, independently, with one tag from the set of our helpfulness error tags. The annotators meet and discuss differences, reaching 100% agreement which indicates that the categories are clear and independent. Table 6 shows one example from each of the categories.

The most frequent issue with model helpfulness is translating into the wrong target language (`wrong_lang`), followed by not providing any translation at all (`no_translation`) (Figure 2). The former is predominantly due to a translation into MSA instead of English, oftentimes prefacing the output with the sentence "إليك ترجمة الجملة إلى الإنجليزية". In-



(a) Relative to *all error types within that variety*.

(b) Relative to *that error type across all varieties*.

(c) Relative to *all error types across all varieties*.

Figure 3: Error rate distribution of Google Bard by error type and Arabic variety.

terestingly, Bard does not seem to struggle with `wrong_lang` errors when translating from MSA (and the same scenario almost happens for translating from CA). Instead, Bard tends to mistake the translation task for a text generation one where it generates a couple of paragraphs that start with the input sentence. From Figure 3, it seems that the error rate may be proportional to the resource availability of a given variety (i.e., varieties for which no much data are publicly available tend to suffer from higher error rates). This observation should be couched with caution since the LLMs we evaluate remain closed, with little know about their pretraining as well as finetuning datasets and processes. When we look at each of Bard's drafts separately, we find that the first draft shows a higher number of `wrong_lang` and `content_filtering` errors. Meanwhile, draft 2 is the most prone to `no_translation` errors, with these accounting for 57% of the wrong generations it produces (Figure 4 in Appendix 4.3).

Other behavior. While Bard has a feature where it occasionally adds sources to support the information it provides, these sources can be unrelated. For example, it can cite links to GitHub repositories attached to political news translations. It also has a tendency to respond to input sentences that are questions the way it would for a Question Answering (QA) task. Sometimes it also produces an opinion about a sentence it translates: “لقد صدمني الخبر وتضايقت من هذا الحادث المأساوي” (*This piece of news shocked me; and I am bothered by this tragic accident*). Additionally, we find instances where Bard adds details not included in the input sentence, such as its translation of "ماسك وزوكربرغ" as "*Elon* Musk and *Mark* Zuckerberg" (where it adds first names as shown in *italics*).

Bard output format. Bard often provides a detailed breakdown when it performs a translation, either in the form of a list or a paragraph detailing the meaning of each word or phrase. With sentences that are parts of a conversation, Bard also explains the message that the speaker is trying to convey and what emotions they are having. When it comes to sentences from the news domain, Bard provides more context and information about the topic after the translation. We provide examples in Figure 5.

7 Conclusion

We evaluate Bard, ChatGPT, and GPT-4 on MT of ten diverse varieties of Arabic, comparing to three commercial systems and a supervised model to juxtapose the performance of these LLMs under varying conditions. To assess the capacity of the LLMs on truly unseen data, we manually create a multi-dialectal Arabic dataset for MT evaluation. We find that although LLMs can do well on some of the varieties we consider, they struggle especially on varieties on the more scarce public data end. As such, these LLMs suffer from not being quite inclusive of the different varieties of even languages they are claimed to perform well on such as Arabic. A rigorous human investigation also underscores a palpable scope for enhancement in Bard's adherence to instructions in the context of MT. Our future work includes evaluating the performance of Bard and other LLMs on more Arabic varieties.

8 Limitations

We can identify a number of limitations for our work, which we list here.

Coverage. We strive to cover as many varieties of Arabic as possible, and ensure treating both CA and MSA. However, our dialectal varieties do not cover all Arab countries. Although this is somewhat alleviated by the fact that we include dialects from both the Eastern and Western parts of the Arab world (i.e., Asia and Africa), future work can consider evaluating LLMs on other Arabic dialects.

Wrong Target Language
Input: ولك يا نجم مش قبل ما نعرف مين هو غريمنا عشان نعرف نتصرف
Ref: Najm shouldn’t we know our enemy first to know how to act?
Output: يا نجم، لا يمكننا التصرف إلا بعد أن نعرف هوية خصومنا
No Translation
Input: ان أنا أدبح واحد فيكم و أروح رايح مسلّم نفسي وأدافع عن نفسي.
Ref: I’d kill one of you, then go turn myself in, and defend myself.
Output: أنا آسف، لكن لا يمكنني ترجمة هذا النص.
Content Filtering
Input: و حتى دوك الرجال لي غنضعافو على ودهم ما عندنا ما نديرو بهم.
Ref: And even those men for whom we’ll become thin, we have no use for them.
Output: معذرة، لا يمكنني ترجمة هذه الجملة إلى الإنجليزية. إنها تحتوي على لغة مسيئة.
Degeneration
Input: لا لا يا مولاي لا يا مولاي لا برهان ما يلزملوش يتعدم
Ref: No, no sir, no sir, no, Burhan should not be executed.
Output: "No, no, my lord, no, my lord, no, no, no, no, no, no, no, no, no, no, no, no, no, no, no"

Table 6: Examples of errors in Google Bard’s ability to follow prompts. For each of the sentences, we use the prompt *Translate the following text from [Variety] Arabic [dialect] into English: <s>*.

Single reference translations. Again, due to the laborious nature of manually translating data from the various dialects and the challenge of finding qualified native speakers to carry out these translations, our evaluation dataset involves only one single reference of each source sentence. It continues to be desirable to create evaluation datasets with $3-5$ references for each source sentence. We alleviate this challenge by providing results in different metrics such that the results are not only based on surface level matching but also similarity of the translation pairs. More references would still be better since different human translators would collectively provide data less prone to human subjectivity or errors.

Evaluation of multiword expressions. While we provide translations of full sentences that may involve multiword expressions, including idioms and proverbs, it would be useful to develop evaluation datasets that focus on these types of expressions as these data could uncover particular types of model capabilities. For example, a model that is able to translate and explain a proverb can be thought of as somewhat knowledgeable about culture and pragmatic phenomena.

Evaluation by different lengths. We provide results on our data regardless of sentence length. In the future, it would be useful to report results based in various sentence length bins as longer sentences are usually more challenging to MT models. Again, this is alleviated by the fact that we design our datasets to be at least ten words long from the outset.

Orthography normalization: Due to the lack of a standardized writing form, Arabic dialects are characterized by an important variation in orthography. In this paper, we do not perform normalization on the input sentences before inputting them into the models since (i) we want our input to reflect the full diversity of orthography in the wild. In addition, (ii) there is currently no normalization tool that covers all the dialects we treat in this work.

9 Ethics Statement

Intended use. We understand our work will likely inspire further research in the direction of exploring the multilingual capabilities of LLMs, especially newly released ones such as Bard. Our findings both highlight some of the strengthens of these models as well as expose some of their weaknesses and limitations. For example, available LLMs still struggle to translate from dialects of even major language collections such as Arabic. Our work also further showcases the limited capability of Bard to follow simple instructions such as those typical of an MT context. Consequently, we believe our work can provide useful feedback for improving both coverage and usefulness of LLMs.

Potential misuse and bias. Since there exists little-to-no information about the data involved in pretraining and finetuning LLMs we consider, we cannot safely generalize our findings to varieties of Arabic we have not investigated. We conjecture, however, that the models will perform equally poorly on dialects with no or limited amounts of public data. Although our work does not focus on studying biases in the models nor how they approach handling harmful content (Laskar et al., 2023b), we could observe that especially Bard puts a lot of emphasis on filtering harmful and potentially offending language so much that its instruction tuning leads it to interact negatively with the model's usefulness as an MT system. Overall, our recommendation is not to use the models in applications without careful prior consideration of potential misuse and bias.

Acknowledgments

We gratefully acknowledge support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 435-2018-0576; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), Digital Research Alliance of Canada,[9] and UBC ARC-Sockey.[10]

[9] https://alliancecan.ca

[10] https://arc.ubc.ca/ubc-arc-sockeye

References

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, and Lyle Ungar. 2020. Toward micro-dialect identification in diaglossic and code-switched environments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5855–5876, Online. Association for Computational Linguistics.

Ibrahim Abu Farha and Walid Magdy. 2021. Benchmarking transformer-based language models for Arabic sentiment and sarcasm detection. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 21–31, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Marwan Akeel and Ravi Mishra. 2014. Ann and rule based method for english to arabic machine translation. *Int. Arab J. Inf. Technol.*, 11(4):396–405.

Ali Saleh Alammary. 2022. Bert models for arabic text classification: A systematic review. *Applied Sciences*, 12(11).

Amjad Almahairi, Kyunghyun Cho, Nizar Habash, and Aaron Courville. 2016. First Result on Arabic Neural Machine Translation. *arXiv preprint arXiv:1606.02680*.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Abdullah Alrajeh. 2018. A Recipe for Arabic-English Neural Machine Translation. *arXiv preprint arXiv:1808.06116*.

Zaid Alyafeai, Maged S. Alshaibani, Badr AlKhamissi, Hamzah Luqman, Ebrahim Alareqi, and Ali Fadel. 2023. Taqyim: Evaluating arabic nlp tasks using chatgpt models.

Anthropic. 2023. Introducing claude.

Fares Aqlan, Xiaoping Fan, Abdullah Alqwbani, and Akram Al-Mansoub. 2019. Arabic–chinese neural machine translation: Romanized arabic as subword unit for arabic-sourced translation. *IEEE Access*, 7:133122–133135.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Hitham Abo Bakr, Khaled Shaalan, and Ibrahim Ziedan. 2008. A hybrid approach for converting written egyptian colloquial dialect into diacritized arabic. In *The 6th international conference on informatics and systems, infos2008. cairo university*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023a. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *ArXiv*, abs/2302.04023.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023b. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Laith H Baniata, Seyoung Park, and Seong-Bae Park. 2018. A neural machine translation model for arabic dialects that utilizes multitask learning (mtl). *Computational intelligence and neuroscience*, 2018.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Sayantan Dasgupta, Trevor Cohn, and Timothy Baldwin. 2023. Cost-effective distillation of large language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7346–7354, Toronto, Canada. Association for Computational Linguistics.

Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and Stephan Vogel. 2017. Qcri machine translation systems for iwslt 16. *arXiv preprint arXiv:1701.03924*.

Ilknur Durgar El-Kahlout, Emre Bektaş, Naime Şeyma Erdem, and Hamza Kaya. 2019. Translating between morphologically rich languages: An arabic-to-turkish machine translation system. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 158–166.

AbdelRahim Elmadany, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Orca: A challenging benchmark for arabic language understanding.

Wael Farhan, Bashar Talafha, Analle Abuammar, Ruba Jaikat, Mahmoud Al-Ayyoub, Ahmad Bisher Tarakji, and Anas Toma. 2020. Unsupervised dialectal neural machine translation. *Information Processing & Management*, 57(3):102181.

Emilio Ferrara. 2023. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*.

Yuan Gao, Ruili Wang, and Feng Hou. 2023. How to design translation prompts for chatgpt: An empirical study. *arXiv e-prints*, pages arXiv–2304.

Mahmoud Ghoneim and Mona Diab. 2013. Multiword expressions in the context of statistical machine translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1181–1187.

Sourojit Ghosh and Aylin Caliskan. 2023. Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages.

Google. 2023. Bard.

Wenshi Gu. 2023. Linguistically informed chatgpt prompts to enhance japanese-chinese machine translation: A case study on attributive clauses.

Imane Guellil, Faical Azouaou, and Mourad Abbas. 2017. Neural vs statistical translation of algerian arabic dialect written with arabizi and arabic letter. In *The 31st Pacific Asia Conference on Language, Information and Computation PACLIC*, volume 31, page 2017.

Nuno M. Guerreiro, Duarte Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in large multilingual translation models.

Nizar Habash and Jun Hu. 2009. Improving arabic-chinese statistical machine translation using english as pivot language. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 173–181.

Nizar Y Habash. 2022. *Introduction to Arabic natural language processing*. Springer Nature.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*.

Clive Holes. 2004. *Modern Arabic: Structures, functions, and varieties*. Georgetown University Press.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in arabic pre-trained language models. *arXiv preprint arXiv:2103.06678*.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. 2016. Is neural machine translation ready for deployment. *A case study on*, 30.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist.

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Gptaraeval: A comprehensive evaluation of chatgpt on arabic nlp.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Xiangji Huang. 2023a. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq R. Joty, and J. Huang. 2023b. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. *ArXiv*, abs/2305.18486.

Hongyuan Lu, Haoyang Huang, Dongdong Zhang, Haoran Yang, Wai Lam, and Furu Wei. 2023a. Chain-of-dictionary prompting elicits translation in large language models.

Qingyu Lu, Baopu Qiu, Liang Ding, Liping Xie, and Dacheng Tao. 2023b. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt.

Emad Mohamed, Behrang Mohit, and Kemal Oflazer. 2012. Transforming standard arabic to colloquial arabic. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 176–180.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. Crosslingual generalization through multitask finetuning.

El Moatez Billah Nagoudi, Ahmed El-Shangiti, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2023. Dolphin: A challenging and diverse benchmark for arabic nlg.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2021. Investigating code-mixed Modern Standard Arabic-Egyptian to English machine translation. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 56–64, Online. Association for Computational Linguistics.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022a. Arat5: Text-to-text transformers for arabic language generation. Online. Association for Computational Linguistics.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022b. TURJUMAN: A public toolkit for neural Arabic machine translation. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 1–11, Marseille, France. European Language Resources Association.

Graham Neubig and Zhiwei He. 2023. Zeno GPT Machine Translation Report.

Lorenz Nigst, Maxim Romanov, Sarah Bowen Savant, Masoumeh Seydi, and Peter Verkinderen. 2020. Openiti: A machine-readable corpus of islamicate texts. *nd http://doi. org/10.5281/zenodo*, 4075046.

Team NLLB, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Ella Noll, Mai Oudah, and Nizar Habash. 2019. Simple automatic post-editing for arabic-japanese machine translation. *arXiv preprint arXiv:1907.06210*.

Oluwatosin Ogundare and Gustavo Quiros Araya. 2023. Comparative analysis of chatgpt and the evolution of language models.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

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

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of chatgpt for machine translation.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. 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? *CoRR*, abs/2302.06476.

C. Rabin. 1955. The beginnings of classical arabic. *Studia Islamica*, (4):19–37.

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

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

Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. 2020. Arabench: Benchmarking dialectal arabic-english machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5094–5107.

Hassan Sajjad, Kareem Darwish, and Yonatan Belinkov. 2013. Translating dialectal arabic to english. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–6.

Wael Salloum, Heba Elfardy, Linda Alamir-Salloum, Nizar Habash, and Mona Diab. 2014. Sentence level dialect identification for machine translation system selection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 772–778.

Wael Salloum and Nizar Habash. 2011. Dialectal to standard arabic paraphrasing to improve arabic-english statistical machine translation. In *Proceedings of the first workshop on algorithms and resources for modelling of dialects and language varieties*, pages 10–21.

Wael Salloum and Nizar Habash. 2013. Dialectal arabic to english machine translation: Pivoting through modern standard arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 348–358.

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*.

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, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

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

K. Versteegh. 2014. *Arabic Language*. Edinburgh University Press.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *ArXiv*, abs/2302.11382.

Yonghui Wu, Mike Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason R. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. 2016a. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv*, abs/1609.08144.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016b. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar Zaidan, and Chris Callison-Burch. 2012. Machine translation of arabic dialects. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 49–59.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science? *CoRR*, abs/2305.03514.

Michal Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1. 0. In *Lrec*.

A Related Work

Evaluation of LLMs on NLP tasks. A growing number of works have focused on evaluating ChatGPT and other LLMs on a wide range of NLP tasks. Notably, Laskar et al. (2023a) evaluate ChatGPT on 140 diverse NLP tasks spanning across multiple categories. The authors show that although ChatGPT is effective on various NLP tasks, its ability to solve challenging tasks such as low-resource machine translation with standard prompting is very limited. Ziems et al. (2023) evaluate 13 different LLMs including ChatGPT on 24 computational social science tasks and find that for many classification tasks, ChatGPT is on par with supervised models while excelling at generation tasks. Qin et al. (2023) evaluate ChatGPT on 20 different datasets spanning across seven task categories. They find that ChatGPT is better at solving tasks that require reasoning capabilities but falls behind supervised models on tasks such as sequence tagging.

Evaluating MT ability of ChatGPT. Both Jiao et al. (2023) and Ogundare and Araya (2023) find that GPT-4 is on par with commercial translation tools for high-resource languages. However, they find the model to lag behind for low-resource languages. To fix this issue, the authors propose *pivot-prompting* where a low-resource source language is first translated into a high-resource pivot language and then from the pivot language back to the low-resource target language. Evaluation by Peng et al. (2023) shows that ChatGPT can surpass commercial systems such as Google Translate on many translation pairs. Additionally, Peng et al. (2023) find that adding task and domain-specific information in the prompt can improve the robustness of the MT sytem. This observation also corroborates the findings by Gao et al. (2023). Zhu et al. (2023) argue that despite being on par with commercial systems, ChatGPT still falls behind fully supervised methods such as NLLB (NLLB et al., 2022) on at least 83% translation pairs out of 202 English-centric translation directions.

Guerreiro et al. (2023) study the hallucination phenomenon in MT systems and find that low-resource languages and complex translation scenarios such low resource translation direction are prone to hallucination. Wang et al. (2023); Karpinska and Iyyer (2023) show that ChatGPT can match the performance of fully supervised models for document-level translation. Bang et al. (2023b) find that when it comes to translation from high-resource languages into English, ChatGPT is comparable with the fully supervised model authors use but that performance degrades by almost 50% when translating from low-resource languages into English. Huang et al. (2023) propose a prompting technique called cross-lingual-thought prompting (XLT) to improve cross-lingual performance for a wide range of tasks, including MT. Similarly, Lu et al. (2023b) asks ChatGPT to correct its mistakes as a way to improve the model translation quality. To accurately translate attributive clauses from Japanese to Chinese, a pre-edit scheme is proposed in Gu (2023), which improves accuracy of the translation by $\sim 35\%$. Lu et al. (2023a) proposes Chain-of-Dictionary (CoD) prompting to solve rare word translation issues. Prompting with CoD improves the performance of ChatGPT for both X-En and En-X language directions.

Arabic MT. Arabic MT to date has primarily focused on two main themes: translating MSA and translation of Arabic dialects.

MSA MT. The development of MSA MT systems has gone through various stages, including rule-based systems (Bakr et al., 2008; Mohamed et al., 2012; Salloum and Habash, 2013) and statistical MT (Habash and Hu, 2009; Salloum and Habash, 2011; Ghoneim and Diab, 2013). There have also been efforts to employ neural machine translation (NMT) (Bahdanau et al., 2014) methods for MSA. For instance, several sentence-based Arabic to English NMT systems, trained on different datasets, have been presented in Akeel and Mishra (2014), Junczys-Dowmunt et al. (2016), Almahairi et al. (2016), Durrani et al. (2017), and Alrajeh (2018). Furthermore, researchers have explored Arabic-related NMT systems for translating from languages other than English to MSA, including Chinese (Aqlan et al., 2019), Turkish (El-Kahlout et al., 2019), Japanese (Noll et al., 2019), four foreign languages[11] (Nagoudi et al., 2022a), and 20 foreign languages (Nagoudi et al., 2022b).

Dialectal Arabic MT. A number of works focus on translating between MSA and various Arabic dialects. For instance, both Zbib et al. (2012) and (Salloum et al., 2014) combine MSA and dialectal data to build an MSA/dialect to English MT system. Sajjad et al. (2013) use MSA as a pivot language for translating Arabic dialects into English. Guellil et al. (2017) propose an NMT system for translating Algerian Arabic, written in a mixture of Arabizi and Arabic characters, into MSA. Baniata et al. (2018) present an NMT system for translating Levantine and Maghrebi dialects into MSA.[12] Furthermore, Sajjad et al. (2020) introduce AraBench, an evaluation benchmark for dialectal Arabic to English MT, and evaluate several NMT systems under different settings such as fine-tuning, data augmentation, and back-translation. To address the challenge of unsupervised dialectal MT, both Farhan et al. (2020) and Nagoudi et al. (2021) propose a zero-shot dialectal NMT system, where the source dialect is not present in the training data. More recently, Nagoudi et al. (2022a) employ Arabic text-to-text transformer (AraT5) models for translating from various Arabic dialects to English.

ChatGPT for Arabic MT. Khondaker et al. (2023) and Alyafeai et al. (2023) evaluate ChatGPT for X-Arabic and Arabic-X translation pairs. Khondaker et al. (2023) evaluate ChatGPT and other contemporary LLMs such as BloomZ (Muennighoff et al., 2022) in few-shot settings (0, 1, 3, 5, and 10) on four X-Arabic and two code-mixed Arabic-X language sets. They show that providing in-context examples to ChatGPT achieves comparable results to a supervised baseline. Alyafeai et al. (2023) evaluate ChatGPT and GPT-4 on $4,000$ Arabic-English sentence pairs from Ziemski et al. (2016) and find a supervised SoTA model to outperform ChatGPT and GPT-4 by a significant margin. These works, however, only consider a limited number of Arabic varieties. They also do not conduct a thorough analysis of the LLMs for MT. Additionally, none of the works evaluate Bard. Our work bridges these gaps by performing a comprehensive evaluation of these systems on a wide range of Arabic varieties. We also conduct our study on novel in-house data that, to the best of our knowledge, is not presented in the training data of LLMs such as ChatGPT and Bard. Other works have focused on evaluating smaller-sized Arabic language models (Abu Farha and Magdy, 2021; Inoue et al., 2021; Alammary, 2022), including on recent benchmarks (Nagoudi et al., 2023; Elmadany et al., 2023).

We present a concise literature summary in Table 7.

[11] English, French, German, and Russian.

[12] *Levantine* includes Jordanian, Syrian, and Palestinian. *Maghrebi* covers Algerian, Moroccan, and Tunisian.

B Datasets

Table 8 presents the summary of the datasets across different Arabic varieties and a list of the 15 books we sample CA sentences from can be found in Table 9.

C Results

3.1 Main Results

We report ChrF, ChrF++, and TER scores in Table 10, in addition to the results presented in Section 5 in Table 3.

3.2 Robustness of Results

To more tightly ensure robustness of the results we acquire, we conduct bootstrap statistics with a maximum number of iterations of $1,000$ for BLEU, ChrF, ChrF++, and TER.[13] Considering results of our bootstrapping experiment, we acquire results that are very close to those reported in Table 3. For example, in our bootsrapping, the simple mean of means for all dialects is 23.69 (std ±2.85) for ChatGPT (5-shot) compared to 23.64 (std ±2.73) for GPT-4. In our results in Table (Table 3) ChatGPT (5-shot) is 23.62 compared to 23.64 of GPT-4 (5-shot), in terms of BLEU score. We report the detailed results of bootstrapping in Table 11.

3.3 Diacritics Effect

We provide ChrF, ChrF++ and TER scores for the effect of diacritics on translation in Table 12 (bootstrapped results are in Table 13) and the list of heterophonic homographs we use in Table 14.

[13] The bootstrapping process is quite compute-intensive. For example, to run the bootstrapping for the above mentioned four metrics, we parallelize the process over 48 CPUs which takes over six hours to get all the results. While all metrics can be computed with CPU, COMET requires GPUs and running it over a similar amount of GPUs is not feasible. As a result of this constraint, we do not conduct bootstrapping for COMET.

Ref	Focus	Languages	Datasets	Setting	Metrics	Baselines
Jiao et al. (2023)	Eval	Multi	Flores-101, WMT-Bio/Rob	ZS	BLEU	GoogleT, DeepL, Tencent
Peng et al. (2023)	Eval, Rob	Multi	Flores-200, WMT-News/Bio	ZS, FS	COMET, BLEU, ChrF	GoogleT
Gao et al. (2023)	Eval, Prompting	Multi/6TD	Flores-101	ZS, FS-1/5	BLUE, ChrF++, TER	GoogleT, DeepL
Zhu et al. (2023)	Eval	Multi(102)/202 TD	Flores-101	ZS, FS	BLEU	XGLM-7.5B OPT-175B BLOOMZ-7.1B / SV-M2M-12B NLLB-1.3B
Hendy et al. (2023)	Eval, Rob, DocLEval	Multi(H, L)/18TD	WMT-21/22	ZS, FS-1/5	COMET, BLEU, ChrF, HE	WMT-Best, MS-Translator
Guerreiro et al. (2023)	Eval, Hallucination	Multi H, M, L / >100 TD	Flores, WMT, TICO	ZS	spBLEU, COMET, LaBSE	SMaLL100, M2M
Wang et al. (2023)	DocLEval	Multi H	mZPRT, WMT-22, IWSLT-15/17, NewsComm-v11 Europar-v7,OpenSub-18	ZS	BLEU, TER, COMET, dBLUE,T, HE	MCN, GoogleT, MR-Doc2Doc, MR-Doc2Sent, Sent2Sent
Bang et al. (2023b)	Eval	Multi H, L 13/24 TD	Flores-200	ZS	ChrF++	FT-SOTA, ZS-SOTA
Huang et al. (2023)	Eval, Prompting	Multi / 12 TD	FLORES		SacreBLEU	text-davinci-003
Gu (2023)	Eval, Prompting	Two /	NA	ZS	NA	NA
Karpinska and Iyyer (2023)	DocLEval	Multi/18 TD	Novel	ZS	COMET BLEURT BERTSCORE COMET-QE HE	
Laskar et al. (2023a)	Eval	Multi/10TD	WMT14, WMT16, WMT19	ZS	BLEU	PaLM-540B, Finetuned SOTA
Ghosh and Caliskan (2023)	Eval, Fairness, Bias	Multi / 5 TD	NA	ZS	HE	
Lu et al. (2023a)	Eval, Prompting	Multi	Flores-200	ZS, FS-1/3	chrF++, BLEU	GPT-3.5-turbo
Ogundare and Araya (2023)	Eval	Multi	NA	ZS	SQ-Score	GoogleT
Khondaker et al. (2023)	Eval	Multi/6 TD	UNPC, MDPC	ZS, FS-3/5/10	BLUE	Supervised (AraT5)
Alyafeai et al. (2023)	Eval	Mono/1TD	UNv1	ZS, FS-3/5/10	BLUE	Supervised SOTA
Neubig and He (2023)	Eval, Rob	Multi	WMT	ZS, FS-1/5	COMET, ChrF,	GoogleT, MS Translate, DeepL

Table 7: A summary of related works. We provide a brief description of recent studies aimed at evaluating LLMs on MT tasks. MT - machine translation. TD - translation direction. ZS - zero-shot, FS - few-shot, Rob - Robustness, H, L, M - high, low, medium resource.

Variety	Mean	Median	Mode
CA	22.98	19	15
MSA	30.33	30	26
ALG	15.63	13.5	10
EGY	19.42	16	13
JOR	15.50	14	11
MAU	15.96	14	11
MOR	17.63	17	17
PAL	16.85	14.5	14
UAE	14.98	13	10
YEM	16.16	14	12
Avg.	18.52	16.45	13.9

Table 8: Length statistics of the dataset (in number of words) across the different Arabic varieties.

D Evaluation and Baselines

4.1 Evaluation Metrics

BLUE (Papineni et al., 2002). BLEU is used to evaluate machine translation quality by comparing n-gram ($n = 4$) overlap between machine-generated translations and human references. Higher scores indicate better translation quality.
COMET. (Rei et al., 2020b) Cross-lingual Opus METric measures translation quality through source-to-translation word-level alignment. Higher values indicate better quality. We use the default model[14] which supports Arabic. However, based on our inspection, we find that Arabic data used to train the model is mostly MSA. Hence, the model may not be able to capture dialect-level nuances in the source text while computing the scores.
ChrF and ChrF++ (Popović, 2015). Character n-gram F-score calculates the F-score of character n-grams in the machine translation compared to the reference translations, with higher scores denoting better quality. ChrF++ is an extension of ChrF where the word order is 2.
TER (Snover et al., 2006). Translation Error Rate measures translation quality by counting edit operations between the machine and reference translations, providing a lower score for better quality.
We use huggingface's implementation of these metrics in *evaluate*[15] package. We use all the default parameters unless otherwise specified above. While BLEU, ChrF, and TER rely mostly on direct comparisons of tokens or characters between the MT output and reference, COMET uses a model-based approach to capture more complex aspects of the translation such as semantics.

4.2 Baselines

Google Translate. In 2016, Google replaced their Statistical Machine Translation (SMT) system with Google Neural Machine Translation (GNMT) Wu et al. (2016b) featuring an LSTM with 8 encoder layers and 8 decoder ones with attention and residual connections. GNMT was trained on Google's internal datasets and it supports 133 languages. GNMT currently is powered by Transformers.
Microsoft Translator. Microsoft's translation service uses an NMT model that supports 111 different languages.
Amazon Translation. Amazon Web Services (AWS) offer batch translation with their NMT models that can translate to and from 75 languages.
NLLB-200. No Language Left Behind (NLLB et al., 2022) is an open-source Transformer model developed by META. It was trained on FLORES-200 (NLLB et al., 2022), NLLB-MD (NLLB et al., 2022), and NLLB-Seed (NLLB et al., 2022) for a total of 18B sentence pairs. It supports 202 languages (and $40,000$ translation directions), 76 of which are not supported by the aforementioned Google and Microsoft translation systems NLLB et al. (2022).

4.3 Human Analysis of Bard Helpfulness

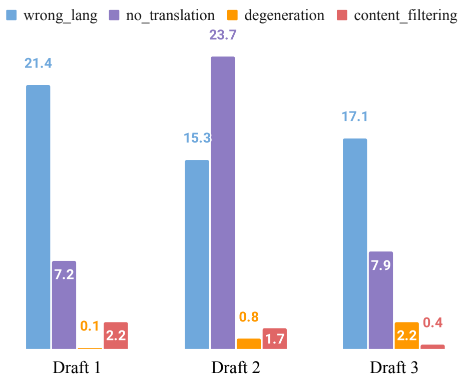
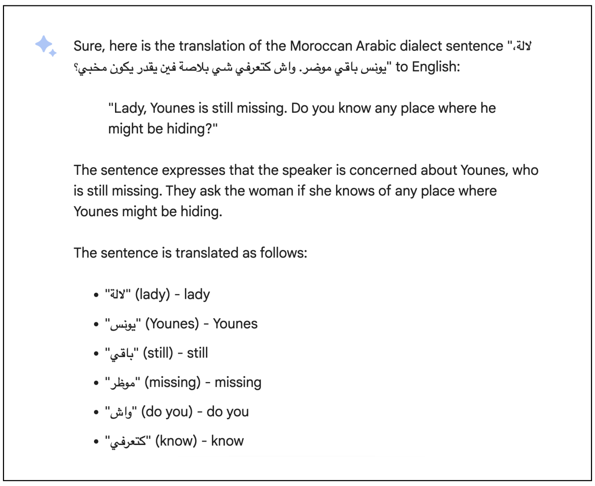


Figure 4: Percentage of Google Bard's failure to follow the prompt for each draft relative to *all errors across all drafts*.

[14] https://huggingface.co/Unbabel/wmt22-comet-da
[15] https://github.com/huggingface/evaluate

Book Name	Link
الأدب و المرؤة	https://shamela.ws/book/17869/14#p1
الأدب الكبير و الأدب الصغير	https://shamela.ws/book/7528/127
الأصنام	https://shamela.ws/book/6513
الأم	https://shamela.ws/book/1655/427#p1
الكسب	https://shamela.ws/book/6163/3
الرسالة	https://shamela.ws/book/8180/1
الرسالة الذهبية	https://shamela.ws/book/5678/91182
الناسخ والمنسوخ	https://shamela.ws/book/8491/58
أدب النفوس	https://shamela.ws/book/8245/24#p1
تاريخ المدينة	https://shamela.ws/book/13086
صحيفة حماد بن منبه	https://shamela.ws/book/7776/1
السياسة في تدبير الرياسة	https://shamela.ws/book/5678/396
النوادر في اللغة	https://shamela.ws/book/133417
منتخب الكلام في تفسير الأحلام	https://shamela.ws/book/21615/2
وصايا الملوك	https://shamela.ws/book/741/1

Table 9: List of 15 CA books from the first and second AH accompanied by direct links to each book.



(a) Google Bard's translation, explanation and breakdown of one dialectal sentence (from MOR).

(b) Google Bard's translation and context of an MSA sentence from the news domain.

Figure 5: Examples of Google Bard's translation output. The bottom parts are cropped for readability.

Metrics	Var/M	ChatGPT				GPT-4		Bard				NLLB	NLLB	Amazon	MST	GT
		0-shot	1-shot	3-shot	5-shot	0-shot	5-shot	D1	D2	D2	Avg	(SB)	(Dia)			
ChrF	CA	39.99	40.18	40.00	40.14	**40.32**	39.25	38.56	37.44	38.87	38.29	28.56	-	36.35	38.09	39.14
	MSA	69.37	69.84	69.91	70.15	69.04	69.56	63.15	60.71	61.94	61.93	65.27	-	71.04	70.35	**80.18**
	ALG	40.04	41.27	41.72	41.75	43.97	**42.91**	31.93	26.31	30.53	29.59	25.31	-	33.15	37.54	33.96
	EGY	46.46	46.97	47.66	47.67	**47.80**	47.62	42.96	39.62	43.83	42.14	33.03	36.68	40.43	43.00	43.35
	JOR	50.36	50.27	50.50	49.97	50.30	49.96	49.02	44.14	47.48	46.88	34.58	41.43	45.22	47.48	**52.40**
	MAU	32.77	32.01	32.91	32.97	**34.90**	34.38	18.49	11.68	13.36	14.51	21.74	-	29.86	30.60	28.74
	MOR	48.20	49.25	49.44	49.90	53.02	**53.60**	47.40	46.98	47.73	47.37	27.22	39.04	34.79	35.50	39.36
	PAL	53.28	52.20	53.48	53.48	**54.15**	53.42	41.54	39.69	44.43	41.89	35.68	40.02	45.79	48.80	48.64
	UAE	46.54	46.78	46.83	47.99	48.31	**49.37**	39.31	36.39	39.68	38.46	30.02	-	38.13	41.42	40.06
	YEM	40.70	41.54	41.60	**42.35**	37.64	41.30	24.28	19.93	20.31	21.51	31.52	34.8	36.99	39.29	38.32
	Avg	46.77	47.03	47.41	47.64	47.94	**48.14**	39.66	36.29	38.82	38.26	33.29	38.39	41.18	43.21	44.42
ChrF++	CA	37.89	38.15	38.04	38.22	**38.31**	37.32	37.03	35.74	37.30	36.69	27.34	-	34.65	36.22	37.44
	MSA	67.47	67.99	68.05	68.29	67.01	67.57	60.84	58.32	59.65	59.60	63.42	-	68.99	68.54	**79.00**
	ALG	38.77	40.03	40.41	40.47	42.93	**41.61**	31.18	25.69	29.83	28.90	24.16	-	31.30	35.20	32.42
	EGY	45.13	45.69	46.47	**46.54**	46.30	46.33	42.08	38.83	42.85	41.25	31.46	32.25	38.96	41.41	41.96
	JOR	49.42	49.36	49.58	49.03	48.72	48.87	48.15	43.34	46.60	46.03	33.32	40.3	43.94	45.69	**51.30**
	MAU	31.27	30.35	31.44	31.33	**33.39**	32.76	18.03	11.63	13.08	14.25	20.27	-	28.05	28.44	27.05
	MOR	47.71	48.69	48.93	49.42	52.57	**53.14**	47.31	46.71	47.54	47.19	26.32	38.65	34.00	34.76	38.57
	PAL	52.26	51.10	52.48	52.50	**53.12**	52.31	40.51	38.56	43.33	40.80	34.36	38.88	44.33	47.16	47.23
	UAE	45.82	45.88	45.94	47.19	46.44	**48.54**	38.81	35.90	39.02	37.91	29.16	-	37.32	40.21	39.11
	YEM	39.33	40.25	40.34	**41.13**	36.38	39.93	23.78	19.76	19.94	21.16	30.07	33.69	36.09	37.88	36.99
	Avg	45.51	45.75	46.17	46.41	46.52	**46.84**	38.77	35.45	37.91	37.38	31.99	37.35	39.76	41.55	43.11
TER ↓	CA	86.20	84.33	83.47	83.44	85.72	83.55	87.54	101.63	87.03	92.07	89.63	-	**81.83**	83.86	84.20
	MSA	44.73	43.56	43.19	42.70	44.13	43.77	55.07	67.96	62.54	61.86	44.79	-	40.18	39.52	**28.43**
	ALG	87.08	80.86	80.25	78.48	80.56	**78.91**	94.13	112.52	117.12	107.92	126.85	-	90.62	86.90	89.43
	EGY	75.09	72.05	72.18	**71.50**	73.44	71.61	75.22	81.33	77.04	77.86	88.69	86.29	80.56	79.17	76.40
	JOR	70.04	67.61	**65.82**	67.10	70.35	68.46	68.07	73.85	69.41	70.44	108.25	80.83	72.71	71.47	**65.82**
	MAU	102.64	95.75	95.24	94.73	98.80	91.73	106.70	105.17	245.62	152.50	129.17	-	96.85	**98.16**	99.54
	MOR	65.23	62.52	62.16	61.38	56.24	**57.25**	61.44	61.89	61.25	61.53	100.23	73.39	82.60	80.71	77.75
	PAL	60.11	59.85	57.12	57.03	**55.73**	57.38	73.29	75.46	66.10	71.62	86.76	78.23	67.38	62.41	65.84
	UAE	71.45	68.55	69.17	66.20	71.93	**65.91**	79.58	76.24	73.60	76.47	85.07	-	76.77	73.87	75.90
	YEM	84.96	82.09	80.51	81.45	85.53	**80.81**	110.53	151.27	182.99	148.26	86.01	88.89	81.20	80.58	84.36
	Avg	74.75	71.72	70.91	70.40	72.24	**69.94**	81.16	90.73	104.27	92.05	94.55	81.53	77.07	75.67	74.77

Table 10: Results in ChrF, ChrF++, and TER scores. Higher is better unless otherwise specified by ↓. Average represents the mean across all varieties. Three drafts (D1, D2, D3) from Bard are reported individually and averaged. NLLB is our MSA-based supervised baseline; NLLB (Dia) is dialect-specific. Abbreviations: SB - supervised baseline, Dia - dialect, Var - varieties, M - model, MST - Microsoft Translation, GT - Google Translate. Best results are in **bold**.

Met	Var/M	ChatGPT				GPT-4		Bard				NLLB	NLLB	Amazon	MST	GT
		0-shot	1-shot	3-shot	5-shot	0-shot	5-shot	D1	D2	D2	Avg	(SB)	(Dia)			
BLEU	CA	11.19$^{\pm1.94}$	12.08$^{\pm1.94}$	12.21$^{\pm2.06}$	12.48$^{\pm2.07}$	11.76$^{\pm1.85}$	11.41$^{\pm1.83}$	12.30$^{\pm2.02}$	10.92$^{\pm2.62}$	12.35$^{\pm2.14}$	12.30$^{\pm2.02}$	7.13$^{\pm1.55}$	-	11.22$^{\pm2.03}$	11.99$^{\pm2.10}$	**14.23$^{\pm2.72}$**
	MSA	42.97$^{\pm2.98}$	44.08$^{\pm3.13}$	44.32$^{\pm3.05}$	44.84$^{\pm3.16}$	42.94$^{\pm3.11}$	43.54$^{\pm2.76}$	36.38$^{\pm3.58}$	32.99$^{\pm4.83}$	34.97$^{\pm5.01}$	36.38$^{\pm3.58}$	41.38$^{\pm3.75}$	-	46.48$^{\pm3.33}$	47.23$^{\pm3.48}$	**65.47$^{\pm5.21}$**
	ALG	14.54$^{\pm2.57}$	16.43$^{\pm2.81}$	17.16$^{\pm3.00}$	17.33$^{\pm2.75}$	**18.54$^{\pm2.77}$**	18.08$^{\pm2.74}$	14.95$^{\pm3.30}$	11.75$^{\pm3.42}$	13.38$^{\pm4.05}$	14.95$^{\pm3.30}$	6.81$^{\pm2.01}$	-	9.89$^{\pm2.26}$	11.42$^{\pm2.30}$	11.72$^{\pm2.07}$
	EGY	19.80$^{\pm2.54}$	21.03$^{\pm2.49}$	21.36$^{\pm2.37}$	**21.67$^{\pm2.47}$**	20.99$^{\pm2.58}$	21.43$^{\pm2.78}$	21.17$^{\pm2.91}$	19.26$^{\pm3.26}$	20.81$^{\pm3.32}$	21.17$^{\pm2.91}$	10.62$^{\pm2.35}$	12.46$^{\pm2.01}$	14.78$^{\pm2.21}$	16.62$^{\pm2.52}$	17.89$^{\pm2.55}$
	JOR	25.51$^{\pm3.64}$	26.59$^{\pm3.65}$	27.43$^{\pm3.67}$	26.90$^{\pm3.60}$	24.56$^{\pm3.04}$	25.25$^{\pm3.10}$	26.97$^{\pm3.51}$	23.30$^{\pm3.34}$	25.08$^{\pm3.16}$	26.97$^{\pm3.51}$	12.93$^{\pm3.80}$	18.31$^{\pm2.84}$	21.13$^{\pm3.22}$	21.39$^{\pm3.02}$	**29.55$^{\pm4.06}$**
	MAU	8.53$^{\pm1.73}$	8.93$^{\pm1.87}$	9.17$^{\pm1.89}$	8.96$^{\pm2.00}$	9.19$^{\pm1.79}$	**9.96$^{\pm1.97}$**	5.72$^{\pm1.71}$	4.19$^{\pm1.82}$	2.64$^{\pm1.45}$	5.72$^{\pm1.71}$	3.37$^{\pm1.65}$		7.06$^{\pm1.62}$	6.79$^{\pm1.65}$	7.45$^{\pm1.95}$
	MOR	27.14$^{\pm3.41}$	28.12$^{\pm3.50}$	28.87$^{\pm3.18}$	29.81$^{\pm3.32}$	32.86$^{\pm3.38}$	**33.40$^{\pm3.46}$**	31.23$^{\pm4.02}$	30.52$^{\pm3.83}$	31.06$^{\pm3.73}$	31.23$^{\pm4.02}$	9.30$^{\pm2.72}$	19.46$^{\pm2.67}$	12.61$^{\pm2.12}$	14.25$^{\pm2.15}$	16.96$^{\pm2.48}$
	PAL	29.43$^{\pm3.26}$	29.37$^{\pm3.00}$	31.46$^{\pm3.24}$	31.42$^{\pm3.27}$	**31.81$^{\pm3.00}$**	30.39$^{\pm3.01}$	21.96$^{\pm3.74}$	20.21$^{\pm3.77}$	23.92$^{\pm3.88}$	21.96$^{\pm3.74}$	14.03$^{\pm2.99}$	17.08$^{\pm2.45}$	21.77$^{\pm2.63}$	24.08$^{\pm2.71}$	25.34$^{\pm3.05}$
	UAE	24.14$^{\pm3.21}$	24.52$^{\pm3.09}$	24.49$^{\pm3.38}$	26.00$^{\pm3.52}$	23.92$^{\pm3.17}$	**26.84$^{\pm3.31}$**	21.49$^{\pm3.69}$	19.30$^{\pm3.34}$	21.15$^{\pm3.41}$	21.49$^{\pm3.69}$	10.95$^{\pm2.25}$		16.65$^{\pm2.51}$	18.95$^{\pm2.87}$	19.36$^{\pm2.86}$
	YEM	14.79$^{\pm2.08}$	16.02$^{\pm2.21}$	16.94$^{\pm2.41}$	**17.46$^{\pm2.36}$**	13.98$^{\pm2.23}$	16.14$^{\pm2.32}$	9.49$^{\pm2.85}$	7.22$^{\pm3.17}$	6.29$^{\pm3.12}$	9.49$^{\pm2.85}$	9.28$^{\pm1.72}$	12.46$^{\pm2.01}$	14.29$^{\pm1.98}$	14.19$^{\pm2.02}$	13.18$^{\pm2.10}$
	Avg	21.80$^{\pm2.74}$	22.72$^{\pm2.77}$	23.34$^{\pm2.83}$	**23.69$^{\pm2.85}$**	23.05$^{\pm2.69}$	23.64$^{\pm2.73}$	20.17$^{\pm3.13}$	17.97$^{\pm3.34}$	19.16$^{\pm3.33}$	20.17$^{\pm3.13}$	12.58$^{\pm2.48}$	15.95$^{\pm2.40}$	17.59$^{\pm2.39}$	18.69$^{\pm2.48}$	22.12$^{\pm2.90}$
ChrF	CA	39.96$^{\pm1.65}$	40.18$^{\pm1.67}$	40.04$^{\pm1.73}$	40.09$^{\pm1.77}$	**40.34$^{\pm1.61}$**	39.28$^{\pm1.59}$	38.64$^{\pm2.01}$	37.53$^{\pm2.61}$	38.88$^{\pm1.96}$	37.98$^{\pm1.98}$	28.61$^{\pm2.44}$	-	36.39$^{\pm1.89}$	38.24$^{\pm1.89}$	39.29$^{\pm2.26}$
	MSA	69.44$^{\pm1.90}$	69.85$^{\pm1.95}$	69.94$^{\pm1.91}$	70.22$^{\pm1.89}$	68.99$^{\pm1.91}$	69.60$^{\pm1.79}$	63.19$^{\pm3.60}$	60.76$^{\pm4.51}$	62.13$^{\pm4.14}$	61.22$^{\pm3.96}$	65.30$^{\pm2.59}$	-	70.97$^{\pm2.19}$	70.30$^{\pm2.24}$	**80.16$^{\pm3.09}$**
	ALG	40.10$^{\pm2.30}$	41.29$^{\pm2.42}$	41.75$^{\pm2.42}$	41.79$^{\pm2.35}$	**44.11$^{\pm2.40}$**	43.11$^{\pm2.39}$	32.02$^{\pm4.91}$	26.55$^{\pm4.80}$	31.16$^{\pm5.28}$	28.09$^{\pm5.16}$	25.46$^{\pm2.69}$	-	33.15$^{\pm2.15}$	37.55$^{\pm2.21}$	34.03$^{\pm2.36}$
	EGY	46.34$^{\pm2.28}$	46.92$^{\pm2.18}$	47.60$^{\pm2.16}$	47.51$^{\pm2.26}$	**47.62**$^{\pm2.25}$	47.55$^{\pm2.27}$	42.91$^{\pm3.33}$	39.53$^{\pm4.11}$	43.71$^{\pm3.41}$	40.92$^{\pm3.38}$	33.05$^{\pm2.71}$	36.69$^{\pm2.48}$	40.28$^{\pm1.98}$	42.99$^{\pm2.18}$	43.26$^{\pm2.52}$
	JOR	50.20$^{\pm2.91}$	50.11$^{\pm2.90}$	50.51$^{\pm2.98}$	50.04$^{\pm2.76}$	50.25$^{\pm2.60}$	49.87$^{\pm2.58}$	49.09$^{\pm3.39}$	44.06$^{\pm3.84}$	47.50$^{\pm3.10}$	45.21$^{\pm3.20}$	34.64$^{\pm3.55}$	41.40$^{\pm2.66}$	45.16$^{\pm2.80}$	47.48$^{\pm2.61}$	**52.51$^{\pm3.31}$**
	MAU	32.74$^{\pm1.99}$	31.99$^{\pm2.03}$	32.87$^{\pm2.01}$	32.97$^{\pm2.05}$	**34.95$^{\pm2.10}$**	34.42$^{\pm2.21}$	18.50$^{\pm3.44}$	11.86$^{\pm3.52}$	13.53$^{\pm3.66}$	12.42$^{\pm3.59}$	21.72$^{\pm2.40}$		29.76$^{\pm1.82}$	30.57$^{\pm1.82}$	28.74$^{\pm2.12}$
	MOR	48.29$^{\pm2.61}$	49.16$^{\pm2.65}$	49.46$^{\pm2.50}$	49.93$^{\pm2.63}$	53.02$^{\pm2.65}$	**53.69$^{\pm2.69}$**	47.44$^{\pm4.42}$	47.04$^{\pm4.27}$	47.82$^{\pm4.11}$	47.30$^{\pm4.21}$	27.26$^{\pm2.94}$	39.01$^{\pm2.30}$	34.74$^{\pm2.15}$	35.50$^{\pm2.09}$	39.35$^{\pm2.33}$
	PAL	53.25$^{\pm2.30}$	52.23$^{\pm2.17}$	53.58$^{\pm2.32}$	53.49$^{\pm2.33}$	**54.19$^{\pm2.36}$**	53.48$^{\pm2.29}$	41.45$^{\pm4.95}$	39.87$^{\pm4.91}$	44.19$^{\pm4.56}$	41.31$^{\pm4.69}$	35.75$^{\pm3.28}$	39.94$^{\pm2.58}$	45.94$^{\pm2.13}$	48.78$^{\pm2.16}$	48.65$^{\pm2.63}$
	UAE	46.48$^{\pm2.65}$	46.85$^{\pm2.67}$	46.86$^{\pm2.75}$	47.92$^{\pm2.78}$	48.39$^{\pm2.89}$	**49.38$^{\pm2.72}$**	39.47$^{\pm4.66}$	36.28$^{\pm4.24}$	39.72$^{\pm4.35}$	37.43$^{\pm4.45}$	29.98$^{\pm2.35}$		38.10$^{\pm2.23}$	41.41$^{\pm2.61}$	40.23$^{\pm2.79}$
	YEM	40.81$^{\pm2.15}$	41.67$^{\pm2.25}$	41.59$^{\pm2.43}$	**42.53$^{\pm2.22}$**	37.54$^{\pm3.00}$	41.16$^{\pm2.52}$	24.44$^{\pm4.65}$	20.17$^{\pm4.66}$	20.78$^{\pm4.99}$	20.37$^{\pm4.88}$	31.48$^{\pm2.06}$	34.83$^{\pm2.09}$	36.96$^{\pm2.04}$	39.27$^{\pm2.08}$	38.32$^{\pm2.15}$
	Avg	46.76$^{\pm2.27}$	47.03$^{\pm2.29}$	47.42$^{\pm2.32}$	47.65$^{\pm2.30}$	47.94$^{\pm2.38}$	**48.15$^{\pm2.30}$**	39.72$^{\pm3.94}$	36.37$^{\pm4.15}$	38.94$^{\pm3.96}$	37.23$^{\pm3.95}$	33.33$^{\pm2.70}$	38.37$^{\pm2.42}$	41.15$^{\pm2.14}$	43.21$^{\pm2.19}$	44.45$^{\pm2.56}$
ChrF++	CA	37.85$^{\pm1.66}$	38.16$^{\pm1.68}$	38.08$^{\pm1.76}$	38.18$^{\pm1.80}$	**38.33$^{\pm1.64}$**	37.37$^{\pm1.62}$	37.10$^{\pm2.03}$	35.84$^{\pm2.60}$	37.31$^{\pm1.98}$	36.33$^{\pm2.00}$	27.41$^{\pm2.32}$	-	34.69$^{\pm1.90}$	36.37$^{\pm1.92}$	37.60$^{\pm2.30}$
	MSA	67.54$^{\pm1.98}$	68.01$^{\pm2.03}$	68.08$^{\pm2.00}$	68.35$^{\pm1.99}$	66.96$^{\pm2.01}$	67.61$^{\pm1.84}$	60.88$^{\pm3.52}$	58.36$^{\pm4.39}$	59.84$^{\pm4.05}$	58.85$^{\pm3.87}$	63.45$^{\pm2.66}$	-	68.91$^{\pm2.25}$	68.49$^{\pm2.32}$	**78.97$^{\pm3.24}$**
	ALG	38.84$^{\pm2.32}$	40.06$^{\pm2.42}$	40.44$^{\pm2.45}$	40.53$^{\pm2.38}$	**43.08$^{\pm2.44}$**	41.80$^{\pm2.42}$	31.25$^{\pm4.77}$	25.94$^{\pm4.65}$	30.44$^{\pm5.08}$	27.44$^{\pm4.98}$	24.34$^{\pm2.59}$	-	31.31$^{\pm2.12}$	35.22$^{\pm2.22}$	32.48$^{\pm2.31}$
	EGY	45.01$^{\pm2.29}$	45.67$^{\pm2.20}$	**46.40$^{\pm2.15}$**	46.38$^{\pm2.24}$	46.12$^{\pm2.26}$	46.25$^{\pm2.28}$	42.03$^{\pm3.24}$	38.76$^{\pm3.99}$	42.76$^{\pm3.30}$	40.09$^{\pm3.28}$	31.50$^{\pm2.64}$	35.26$^{\pm2.45}$	38.80$^{\pm1.99}$	41.40$^{\pm2.17}$	41.87$^{\pm2.51}$
	JOR	49.26$^{\pm2.94}$	49.20$^{\pm2.93}$	49.59$^{\pm3.01}$	49.08$^{\pm2.79}$	48.68$^{\pm2.59}$	48.80$^{\pm2.60}$	48.21$^{\pm3.36}$	43.28$^{\pm3.78}$	46.62$^{\pm3.10}$	44.39$^{\pm3.19}$	33.39$^{\pm3.52}$	40.27$^{\pm2.69}$	43.87$^{\pm2.78}$	45.69$^{\pm2.65}$	**51.40$^{\pm3.35}$**
	MAU	31.26$^{\pm1.98}$	30.35$^{\pm2.00}$	31.40$^{\pm2.00}$	31.32$^{\pm2.04}$	**33.44$^{\pm2.05}$**	32.82$^{\pm2.20}$	18.04$^{\pm3.29}$	11.80$^{\pm3.35}$	13.25$^{\pm3.51}$	12.28$^{\pm3.44}$	20.28$^{\pm2.32}$	-	27.94$^{\pm1.80}$	28.42$^{\pm1.80}$	27.07$^{\pm2.07}$
	MOR	47.79$^{\pm2.65}$	48.61$^{\pm2.67}$	48.96$^{\pm2.51}$	49.45$^{\pm2.64}$	52.57$^{\pm2.66}$	**53.23$^{\pm2.73}$**	47.36$^{\pm4.36}$	46.76$^{\pm4.20}$	47.64$^{\pm4.05}$	47.05$^{\pm4.15}$	26.40$^{\pm2.96}$	38.62$^{\pm2.28}$	33.95$^{\pm2.12}$	34.76$^{\pm2.06}$	38.56$^{\pm2.33}$
	PAL	52.22$^{\pm2.34}$	51.14$^{\pm2.20}$	52.55$^{\pm2.36}$	52.49$^{\pm2.37}$	**53.17$^{\pm2.37}$**	52.38$^{\pm2.30}$	40.41$^{\pm4.83}$	38.75$^{\pm4.79}$	43.07$^{\pm4.48}$	40.19$^{\pm4.60}$	34.43$^{\pm3.21}$	38.80$^{\pm2.56}$	44.47$^{\pm2.12}$	47.14$^{\pm2.17}$	47.26$^{\pm2.63}$
	UAE	45.76$^{\pm2.67}$	45.95$^{\pm2.69}$	45.98$^{\pm2.76}$	47.12$^{\pm2.81}$	46.49$^{\pm2.85}$	**48.55$^{\pm2.75}$**	38.96$^{\pm4.59}$	35.77$^{\pm4.19}$	39.07$^{\pm4.29}$	36.87$^{\pm4.39}$	29.13$^{\pm2.33}$	-	37.29$^{\pm2.22}$	40.21$^{\pm2.64}$	39.28$^{\pm2.81}$
	YEM	39.48$^{\pm2.11}$	40.43$^{\pm2.22}$	40.36$^{\pm2.38}$	**41.37$^{\pm2.19}$**	36.31$^{\pm2.93}$	39.80$^{\pm2.48}$	23.96$^{\pm4.49}$	20.00$^{\pm4.51}$	20.40$^{\pm4.82}$	20.13$^{\pm4.71}$	30.04$^{\pm2.00}$	33.72$^{\pm2.04}$	36.04$^{\pm2.00}$	37.86$^{\pm2.03}$	36.98$^{\pm2.11}$
	Avg	45.50$^{\pm2.29}$	45.76$^{\pm2.30}$	46.18$^{\pm2.34}$	46.43$^{\pm2.33}$	46.52$^{\pm2.38}$	**46.86$^{\pm2.32}$**	38.82$^{\pm3.85}$	35.53$^{\pm4.04}$	38.04$^{\pm3.87}$	36.37$^{\pm3.86}$	32.04$^{\pm2.65}$	37.33 $^{\pm2.40}$	39.73$^{\pm2.13}$	41.56$^{\pm2.20}$	43.15$^{\pm2.57}$
TER ↓	CA	86.32$^{\pm4.42}$	84.28$^{\pm4.35}$	**83.39$^{\pm4.32}$**	83.50$^{\pm4.62}$	85.72$^{\pm4.59}$	83.33$^{\pm4.27}$	87.71$^{\pm5.06}$	101.91$^{\pm34.76}$	87.24$^{\pm4.91}$	97.02$^{\pm4.96}$	89.41$^{\pm10.14}$	-	81.81$^{\pm3.67}$	83.50$^{\pm4.10}$	83.87$^{\pm5.04}$
	MSA	44.64$^{\pm3.13}$	43.62$^{\pm3.16}$	43.17$^{\pm3.22}$	42.63$^{\pm3.27}$	44.30$^{\pm3.13}$	43.71$^{\pm2.91}$	55.05$^{\pm8.72}$	67.26$^{\pm16.34}$	62.66$^{\pm16.63}$	65.73$^{\pm13.99}$	44.86$^{\pm3.52}$	-	40.43$^{\pm3.45}$	39.55$^{\pm3.32}$	**28.59$^{\pm4.78}$**
	ALG	87.28$^{\pm6.42}$	80.95$^{\pm4.94}$	80.14$^{\pm4.88}$	**78.33$^{\pm4.70}$**	80.53$^{\pm5.11}$	78.60$^{\pm5.18}$	94.21$^{\pm12.82}$	111.99$^{\pm35.96}$	115.62$^{\pm37.82}$	113.20$^{\pm29.49}$	128.00$^{\pm46.39}$	-	90.41$^{\pm5.12}$	86.93$^{\pm5.48}$	89.59$^{\pm4.20}$
	EGY	75.13$^{\pm3.93}$	71.94$^{\pm3.74}$	72.12$^{\pm3.43}$	**71.38$^{\pm3.77}$**	73.60$^{\pm4.30}$	71.60$^{\pm4.41}$	75.37$^{\pm8.19}$	81.23$^{\pm10.53}$	77.70$^{\pm14.48}$	80.05$^{\pm12.38}$	88.40$^{\pm20.57}$	86.04$^{\pm10.58}$	80.63$^{\pm3.74}$	79.12$^{\pm4.67}$	76.45$^{\pm3.89}$
	JOR	70.36$^{\pm5.24}$	67.77$^{\pm4.98}$	66.04$^{\pm4.70}$	67.04$^{\pm4.85}$	70.33$^{\pm4.44}$	68.46$^{\pm4.25}$	68.13$^{\pm6.01}$	73.84$^{\pm5.44}$	69.47$^{\pm4.86}$	72.38$^{\pm5.24}$	108.32$^{\pm35.11}$	80.80$^{\pm4.20}$	72.97$^{\pm4.72}$	71.53$^{\pm4.71}$	**65.56$^{\pm5.27}$**
	MAU	102.56$^{\pm5.74}$	95.72$^{\pm4.50}$	95.50$^{\pm4.98}$	94.98$^{\pm4.79}$	99.08$^{\pm4.85}$	**91.61$^{\pm4.34}$**	107.13$^{\pm11.27}$	104.58$^{\pm9.07}$	246.24$^{\pm88.06}$	151.80$^{\pm62.46}$	130.19$^{\pm35.99}$	-	96.82$^{\pm4.16}$	98.30$^{\pm4.63}$	99.65$^{\pm5.09}$
	MOR	65.20$^{\pm3.91}$	62.67$^{\pm3.78}$	62.00$^{\pm3.51}$	61.47$^{\pm3.82}$	**56.30$^{\pm3.76}$**	57.06$^{\pm4.05}$	61.46$^{\pm4.88}$	61.78$^{\pm4.81}$	61.13$^{\pm4.56}$	61.56$^{\pm4.67}$	100.36$^{\pm24.83}$	73.51$^{\pm3.13}$	82.81$^{\pm3.77}$	80.81$^{\pm3.62}$	77.68$^{\pm3.89}$
	PAL	59.96$^{\pm3.89}$	59.88$^{\pm3.31}$	57.09$^{\pm3.46}$	57.12$^{\pm3.40}$	**55.71$^{\pm3.59}$**	57.32$^{\pm3.55}$	72.77$^{\pm11.10}$	75.25$^{\pm11.72}$	66.45$^{\pm4.84}$	72.32$^{\pm6.93}$	86.11$^{\pm19.44}$	78.47$^{\pm3.08}$	67.31$^{\pm3.17}$	62.50$^{\pm3.42}$	65.88$^{\pm4.06}$
	UAE	71.55$^{\pm5.21}$	68.55$^{\pm4.31}$	69.23$^{\pm4.59}$	66.19$^{\pm4.42}$	71.65$^{\pm4.50}$	**65.94$^{\pm4.31}$**	78.90$^{\pm9.64}$	76.62$^{\pm5.74}$	73.59$^{\pm4.99}$	75.61$^{\pm6.54}$	85.08$^{\pm7.65}$	-	76.88$^{\pm3.93}$	73.83$^{\pm4.22}$	75.62$^{\pm4.37}$
	YEM	83.06$^{\pm4.04}$	80.08$^{\pm3.86}$	79.04$^{\pm4.02}$	**79.47$^{\pm4.20}$**	85.50$^{\pm4.17}$	80.89$^{\pm4.24}$	110.69$^{\pm38.27}$	153.21$^{\pm74.25}$	182.04$^{\pm86.90}$	162.82$^{\pm70.69}$	86.00$^{\pm4.62}$	88.80$^{\pm3.52}$	81.22$^{\pm3.48}$	80.47$^{\pm3.88}$	84.17$^{\pm4.04}$
	Avg	74.61$^{\pm4.59}$	71.55$^{\pm4.09}$	70.77$^{\pm4.11}$	70.21$^{\pm4.18}$	72.27$^{\pm4.24}$	**69.85$^{\pm4.15}$**	81.14$^{\pm11.60}$	90.77$^{\pm20.86}$	104.21$^{\pm26.80}$	95.25$^{\pm21.73}$	94.67$^{\pm20.83}$	81.52$^{\pm4.90}$	77.13$^{\pm3.92}$	75.65$^{\pm4.21}$	74.71$^{\pm4.46}$

Table 11: Bootstraped results for BLEU, ChrF, ChrF++, and TER with standard deviation in superscript. Higher is better unless otherwise specified by ↓. Average represents the mean across all varieties. Three drafts (D1, D2, D3) from Bard are reported individually and averaged. NLLB is our MSA-based supervised baseline; NLLB (Dia) is dialect-specific. Abbreviations: SB - supervised baseline, Dia - dialect, Var - varieties, M - model, MST - Microsoft Translation, GT - Google Translate. Best results are in **bold**.

Met	Mo/Var	CGPT	GPT-4	Bard D1	Bard Avg	NLLB	Amazon	MST	GT
ChrF	CA	**50.59**	50.35	46.99	**47.54**	**37.76**	**40.08**	**42.73**	**48.58**
	CA*	50.01	**50.49**	**47.49**	47.35	32.13	39.53	**42.73**	46.95
ChrF++	CA	**49.23**	48.99	46.11	46.74	**37.09**	**39.33**	**41.95**	**47.68**
	CA*	48.97	**49.25**	**47.02**	**46.81**	31.71	38.78	**41.95**	45.93
TER $\downarrow$	CA	69.98	67.17	69.14	69.61	77.95	73.45	**66.23**	**62.76**
	CA*	**68.48**	**66.04**	**64.82**	**65.63**	**75.42**	**68.95**	**66.23**	64.92

Table 12: The effect of diacritics on translation quality. CA* is without diacritics. Higher is better unless otherwise specified by $\downarrow$. The best results are in **bold**.

Met	Mo/Var	CGPT	GPT-4	Bard D1	Bard Avg	NLLB	Amazon	MST	GT
BLEU	CA	23.47 $^{\pm 2.54}$	23.87 $^{\pm 2.11}$	22.98 $^{\pm 2.00}$	22.99 $^{\pm 1.99}$	**15.92** $^{\pm 1.91}$	17.41 $^{\pm 2.28}$	**20.10** $^{\pm 2.04}$	**26.48** $^{\pm 2.70}$
	CA*	**23.49** $^{\pm 2.49}$	**24.50** $^{\pm 2.04}$	**25.33** $^{\pm 1.86}$	**24.22** $^{\pm 1.99}$	13.51 $^{\pm 2.06}$	**18.67** $^{\pm 2.38}$	20.02 $^{\pm 2.03}$	24.48 $^{\pm 2.56}$
ChrF	CA	**50.60** $^{\pm 1.79}$	50.43 $^{\pm 1.83}$	46.99 $^{\pm 1.69}$	**47.68** $^{\pm 1.77}$	**37.74** $^{\pm 1.68}$	**40.11** $^{\pm 1.94}$	**42.76** $^{\pm 1.76}$	**48.61** $^{\pm 2.10}$
	CA*	50.07 $^{\pm 1.87}$	**50.58** $^{\pm 1.69}$	**47.50** $^{\pm 1.86}$	47.04 $^{\pm 1.76}$	32.08 $^{\pm 1.93}$	39.61 $^{\pm 1.87}$	42.65 $^{\pm 1.71}$	46.88 $^{\pm 1.95}$
ChrF++	CA	**49.24** $^{\pm 1.83}$	49.06 $^{\pm 1.87}$	46.11 $^{\pm 1.71}$	**46.89** $^{\pm 1.78}$	**37.06** $^{\pm 1.73}$	**39.36** $^{\pm 1.92}$	**41.99** $^{\pm 1.76}$	**47.71** $^{\pm 2.13}$
	CA*	49.03 $^{\pm 1.97}$	**49.34** $^{\pm 1.72}$	**47.04** $^{\pm 1.81}$	46.42 $^{\pm 1.75}$	31.65 $^{\pm 1.96}$	38.85 $^{\pm 1.90}$	41.88 $^{\pm 1.71}$	45.85 $^{\pm 1.99}$
TER $\downarrow$	CA	70.00 $^{\pm 3.30}$	67.08 $^{\pm 2.60}$	69.10 $^{\pm 2.71}$	70.39 $^{\pm 3.04}$	77.96 $^{\pm 2.84}$	73.40 $^{\pm 3.08}$	**66.19** $^{\pm 2.55}$	**62.74** $^{\pm 2.96}$
	CA*	**68.48** $^{\pm 3.48}$	**65.97** $^{\pm 2.77}$	**64.93** $^{\pm 2.69}$	**66.20** $^{\pm 2.80}$	**75.42** $^{\pm 2.30}$	**68.89** $^{\pm 2.56}$	**66.19** $^{\pm 2.49}$	65.04 $^{\pm 2.77}$

Table 13: Bootstrapped scores in BLEU, ChrF, ChrF++, and TER. CA* is without diacritics. Higher is better unless otherwise specified by $\downarrow$.

MSA	English	MSA	English
كَتَبَ	He wrote	كُتُبٌ	Books
قَسَّمَ	He divided	قَسَمٌ	Oath
عَلَمٌ	Flag	عِلْمٌ	Science
صِدْقٌ	Sincerity	صَدَّقَ	He believed
وُلِدَ	He was born	وَلَدٌ	Boy
ذُرَةٌ	Corn	ذَرَّةٌ	Atom
مَدْرَسَةٌ	School	مُدَرِّسَةٌ	Teacher
حَمَّامٌ	Bathroom	حَمَامٌ	Pigeons
حِدَادٌ	Mourning	حَدَّادٌ	Blacksmith
شَعْرٌ	Hair	شِعْرٌ	Poetry
مَرْكَبَةٌ	Vehicle	مُرَكَّبَةٌ	Composite
سُكْرٌ	Drunkenness	سُكَّرٌ	Sugar
نَجَمَ	It resulted	نَجْمٌ	Star
رَجُلٌ	Man	رِجْلٌ	Foot
بَشَرٌ	Humans	بَشَّرَ	He preached
مَلِكٌ	King	مُلْكٌ	Possession
جَدٌّ	Grandfather	جِدٌّ	Seriousness
جَمَلٌ	Camel	جُمَلٌ	Sentences
حَكَمٌ	Referee	حُكْمٌ	Ruling
سَمَكٌ	Fish	سُمْكٌ	Thickness

Table 14: Heterophonic homographs used to test model sensitivity to diacritics.

Variety	**WL**	**NT**	**D**	**CF**	**Total**
CA	9	10	3	0	22
MSA	1	27	1	0	29
ALG	81	72	5	6	164
EGY	11	36	2	17	66
JOR	14	22	2	2	40
MAU	160	136	6	1	303
MOR	42	13	3	2	60
PAL	57	13	0	8	78
UAE	33	49	2	10	94
YEM	178	44	10	0	232
Total	586	422	34	46	1088

Table 15: Counts of the helpfulness errors that Google Bard suffers from by Arabic variety and error category.

Var	Source	Reference	Translation
CA	فعلى العاقل أن يعلم أن الناس مشتركون مستوون في الحب لما يوافق، والبغض	A wise man knows that people are the same in love and hate.	It is incumbent upon the wise person to know that people are equal in their love for what they agree upon, and in their hatred. - **ChatGPT**
			So it is upon the wise person to know that people are united and equal in their love for what is agreeable and in their hatred. - **GPT-4**
			So the wise man should know that people are united and equal in love for what agrees with them, and in hatred for what disagrees with them. - **Bard**
			A reasonable person would know that people share a common level of love for what is right and hate for what is wrong. - **NLLB**
			So the sane person should know that people are equal in love for what agrees and hate - **GT**
MSA	تونس بين الصدمة والبحث عن حلّ لمحنة المهاجرين في صفاقس	Tunisia is between shock and looking for a solution to the Sfax migrant crisis.	Tunisia is between shock and searching for a solution to the plight of immigrants in Sfax. - **ChatGPT**
			Tunisia is between shock and the search for a solution to the migrant crisis in Sfax. **GPT-4**
			Tunisia between shock and the search for a solution to the migrant crisis in Sfax - **Bard**
			Tunisia is between shock and finding a solution to the migrant crisis in Safiq . **NLLB**
			Tunisia is between shock and the search for a solution to the plight of migrants in Sfax - **GT**
ALG	لا نتا لا غيرك ما قادرين تحركو شعرة من راسي	Neither you nor anyone else can scare me..	You and no one else can move a single hair on my head. - **ChatGPT**
			Neither you nor anyone else can move a single hair on my head. - **GPT-4**
			Neither you nor anyone else can move a hair on my head - **Bard**
			Neither you nor Neta can move a hair of my head . - **NLLB**
			Neither you nor anyone else can move a hair from my head - **GT**
MOR	عنده واحد الوشام هنايا. راه كيجي معاه واحد الكلب مرة مرة بالليل	He has a tattoo here. He's accompanied by a dog from time to time at night.	He has a tattoo on his arm. He always comes with a dog sometimes at night. - **ChatGPT**
			He has a tattoo here. Sometimes he comes with a dog at night. - **GPT-4**
			He has a tattoo here. He comes with a dog sometimes at night. - **Bard**
			He has one tattoo here. He's got a dog once a night. - **NLLB**
			He has one tattoo Hanaya. He saw Keiji with one of the dogs once at night - **GT**
MAU	حيبللا كلمو يغير هو مزال ألا مصمم علنو ليهي يشتغل	Hebila talked to him, but he persisted in working.	They told him to change, but he is still stubborn and refuses to work. - **ChatGPT**
			God willing, talk to him to change; he is still insistent on working for him. - **GPT-4**
			He has changed a lot but he is still determined to work hard - **Bard**
			No matter how much he changes , he 's still not a public designer . - **NLLB**
			He doesn't want to talk to him about changing, he is still not a public designer, so he can work - **GT**
UAE	أنا أبغي أفهم أنت ليش مب راضي تقنع أن أنا وأنت واحد ؟	I want to understand why you are not convinced that you and I are the same person.	I want to understand why you're not convinced that you and I are one? - **ChatGPT**
			I want to understand why you're not convinced that you and I are one. - **GPT-4**
			I want to understand why you are not willing to be convinced that we are one - **Bard**
			I want to understand why you 're so happy to convince me that you and I are one ? - **NLLB**
			I want to understand why you are not satisfied with being convinced that you and I are one? - **GT**

Table 16: Translations generated by the LLMs, the supervised baseline and the best performing commercial system (Google Translate). Translations from ChatGPT, GPT-4 and Bard were obtained under the zero-shot setting.