Helpful or Harmful? The Dual Role of Linguistic Features in LLM-Based Dialectal Machine Translation

Abdelhalim Hafedh Dahou [1], Mohamed Amine Cheragui [2]
[1]GESIS – Leibniz-Institute for the Social Sciences, Cologne, Germany
[2]Department of Mathematics and Computer Science, University of Ahmed DRAIA, Adrar, Algeria
abdelhalim.dahou@gesis.org, m_cheragui@univ-adrar.edu.dz

Abstract

Large Language Models (LLMs) have shown promising results in dialectal machine translation, yet the impact of explicit linguistic features remains underexplored. This paper examines whether part-of-speech (POS) tags and diacritization help or hinder LLM-based translation between Algerian dialect (Darija) and Modern Standard Arabic (MSA). Using a linguistically enriched subset of the PADIC dataset, we conduct bidirectional experiments across several frontier and open-weight LLMs, evaluated with automatic metrics and human judgments of adequacy and fluency. Results reveal a dual and asymmetric effect: diacritics can improve adequacy in the MSA → Algerian dialect direction, while POS tags and forced diacritization often introduce noise, especially for Algerian dialect → MSA translation. We further observe a mismatch between traditional overlap-based metrics and human evaluation, suggesting limitations in current evaluation practices. Overall, explicit linguistic augmentation does not consistently benefit LLM-based dialectal translation and must be applied cautiously. The code, prompts, and datasets are available at: FeatureDual-MT Repository

Keywords: LLM, Machine Translation, POS, Diacritization, Algerian dialect, MSA

1. Introduction

Arabic natural language processing poses significant challenges due to the language's rich morphology, flexible word order, and high degree of lexical ambiguity (Habash, 2010). Two linguistic factors are particularly important for addressing these challenges: Part-of-Speech (POS) information and diacritization. POS tagging provides explicit syntactic cues by identifying the grammatical role of words (Petrov et al., 2012), while diacritics encode short vowels and morphological markers that are typically omitted in written Arabic, resulting in substantial lexical and syntactic ambiguity (Habash and Rambow, 2007; Zitouni et al., 2006). In the absence of these cues, a single surface form may correspond to multiple grammatical categories and meanings, complicating automatic language understanding. Consequently, both POS tagging and diacritization have been shown to be effective for improving disambiguation and downstream Arabic NLP tasks, including machine translation (Diab and Hacioglu, 2004; Habash, 2013).

In machine translation, linguistic features such as POS tags and diacritization remain valuable for morphologically rich languages like Arabic. POS information helps models better capture syntactic dependencies, while diacritization reduces ambiguity by providing missing morphological information (Habash, 2019; Zalmout and Habash, 2020). Prior work has shown that incorporating linguistic annotations can complement end-to-end neural models, improving generalization and robustness, particularly in low-resource and dialectal settings (Sennrich and Haddow, 2016; Currey and Heafield, 2020). For Arabic, combining POS-aware and diacritic-sensitive representations has been shown to enhance disambiguation and translation quality (Abdul-Mageed and Bouamor, 2021; Khalifa et al., 2022). Even with transformer-based and large language model (LLM) architectures, explicit linguistic enrichment can improve handling of complex morphology and closely related language varieties (Edman and Sogaard, 2023; Liu and Neubig, 2024).

In this work, we study the impact of incorporating POS tagging and diacritization as auxiliary linguistic features in bidirectional machine translation between Algerian dialect (Darija) and Modern Standard Arabic (MSA) using LLM-based systems. We conduct a systematic evaluation of POS- and diacritic-aware translation settings in both directions, analyzing their effect on ambiguity reduction and translation quality. Our results shed light on the effect of explicit syntactic and morphological cues in translation accuracy and linguistic consistency for dialectal Arabic.

2. Linguistic Differences Between Algerian Dialect and Modern Standard Arabic

MSA serves as the standardized variety used in formal, written, and institutional contexts, whereas Algerian dialect (*Darja*) is the primary medium of everyday communication. Despite their shared historical origin, the two varieties differ substantially at the phonological, morphological, syntactic, and lexical levels. Algerian dialect is characterized by

Proceedings of The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks, pages 66–75
May 11, 2026.

reduced inflectional morphology, distinct syntactic patterns, extensive lexical borrowing, and the absence of a standardized orthography. These structural differences result in limited mutual intelligibility in spoken contexts and have important implications for linguistic analysis and language technology.

2.1. Phonological Differences

Algerian dialect exhibits significant phonological divergence from MSA. Vowel reduction and elision are common, resulting in consonant clusters that are rare in MSA. Additionally, several consonants undergo systematic variation. The MSA phoneme /q/ is frequently realized as (Holes, 2004; Kouloughli, 1999):

- the palatal sound [g / ڤ] is used in several regional varieties. It is attested in northern cities such as Annaba and Sétif, and is especially prevalent in Bedouin dialects. This sound is also widely employed in southern regions of Algeria, including Adrar, Bechar, and Tamanrasset, where it constitutes a stable phonological feature.
- The glottal stop [ʔ / ا] is used in the Algerian dialect spoken in Tlemcen.
- The post-palatal sound [k / ك] represents a distinctive feature of the Algerian dialect and is not attested in other North African dialects. It is primarily used in rural varieties and is also found in several regions and cities, including Kabylia, Jijel, Msirda, and Trara.

Overall, the letter "ق" is not the only case of phonetic variation in Algerian dialect. Other letters, such as: "ج" ,"غ", and "ث" also display different pronunciations across regions, further illustrating the phonological diversity of the Algerian dialect in comparison with MSA.

2.2. Morphological Differences

Although both varieties share a root-and-pattern morphological system, Algerian dialect shows extensive morphological simplification. Case endings, dual forms, and feminine plural which are obligatory in MSA, are absent in Algerian dialect (Holes, 2004), the first and the second person of the singular form are conjugated in the same way in the Algerian dialect. Verbal paradigms are reduced, and tense–aspect distinctions are often expressed analytically rather than through inflection (Sayahi, 2014). Negation provides another clear contrast, as Algerian dialect typically uses a bipartite negation construction (*ma–. . .–š* / ما–. . .–ش), unlike the single negation particle used in MSA (Kouloughli, 1999).

- **MSA:** لا أكتبُ — *lā aktubu*
- **Algerian dialect:** ما نكتبش — *ma n-ktb-š*

2.3. Syntactic Differences

Syntactically, Algerian dialect tends to favor Subject–Verb–Object (SVO) word order, whereas MSA allows both Verb–Subject–Object (VSO) and SVO structures (Versteegh, 2014). Algerian dialect also makes frequent use of resumptive pronouns in relative clauses and relies on discourse particles rather than inflectional markers (Sayahi, 2014). Another important distinction between MSA and Algerian dialect concerns verb–subject agreement and how it is affected by word order. In MSA, verb agreement varies depending on whether the verb comes before or after the subject: it can be either partial or full. In contrast, dialectal Arabic consistently uses full agreement, no matter where the verb appears in the sentence.
Example:

In MSA:

- وصلَ الطلابُ إلى المدرسة
  waáala al-áullābu ilā al-madrasah
  "The students arrived at the school."
  In this case, the verb appears in the singular form despite the plural subject, illustrating partial agreement.
- الطلابُ وصلوا إلى المدرسة
  al-áullābu waáalū ilā al-madrasah
  "The students arrived at the school."
  Here, the verb fully agrees with the plural subject.

In Algerian dialect:

- الطلاب وصلو للمدرسة
  "The students arrived at the school."
  The verb shows full agreement with the subject regardless of word order.

2.4. Lexical Differences

Lexical variation represents one of the most salient differences between Algerian dialect and MSA. Algerian dialect incorporates numerous loanwords from Berber, French, Turkish, Spanish, and Italian as a result of historical contact (Benrabah, 2014) (see Table 1). These borrowings are often integrated into Arabic morphological patterns and may co-occur with Arabic vocabulary within the same utterance (Bouhadiba, 2017).

Loanword	Source	English
كسكسي	Berber *kseksu*	couscous
كوزينة	French *cuisine*	kitchen
باي	Turkish *bey*	governor
ساباط	Spanish *zapato*	shoe
بانيو	Italian *bagno*	bathroom

Table 1: Frequent loanwords in Algerian dialect by source language.

2.5. Orthography and Standardization

Unlike MSA, which has a standardized orthography, Algerian dialect lacks an official writing system (see Table 2) and is primarily oral. When written, particularly in digital contexts, it may appear in Arabic script, Latin script (Arabizi), or mixed forms, resulting in significant spelling variation (Bassiouney, 2009).

Algerian dialect	MSA	English
درك *dark* دروك *drouk*	الآن	now

Table 2: Example of Algerian dialect variants and Arabizi forms corresponding to MSA.

3. Literature review

3.1. Machine Translation in the Era of LLMs

Recent work has shown a clear shift from traditional supervised neural machine translation (NMT) toward LLMs for Dialectal Arabic (DA) to MSA translation. (Elneima et al., 2024) introduced a large-scale benchmark for DA–MSA translation covering five major dialect groups and demonstrated that zero-shot and few-shot prompting with ChatGPT outperforms fine-tuned NMT systems, establishing strong baselines for LLM-based approaches. Similarly, (Atwany et al., 2024) compared supervised Transformer models with prompting-based LLMs and found that few-shot prompting with GPT-3.5 substantially surpasses fine-tuned models trained on MADAR (Bouamor et al., 2018), highlighting the difficulty of dialect modeling using conventional supervision alone.

Several studies explored combining LLMs with data-centric approaches. (Abdelaziz et al., 2024) leveraged ChatGPT-generated translations to construct large-scale parallel corpora, showing that training higher-capacity Transformer models on mixed synthetic and human-curated data significantly improves DA–MSA translation quality. In a similar vein, (Khered et al., 2025) introduced Dial2MSA-Verified, a high-quality multi-dialect dataset with human verification and multiple references, and demonstrated that AraT5 achieves strong performance across dialects, particularly under joint multi-dialect training.

Other work examined efficiency and modeling trade-offs. (Alabdullah et al., 2025) evaluated both training-free prompting and resource-efficient fine-tuning across multiple LLMs, showing that few-shot prompting consistently outperforms zero-shot strategies, while carefully scaled fine-tuning can exceed prompting-based systems when sufficient high-quality data is available. Focusing on Algerian dialect, (Babaali et al., 2025) compared traditional Seq2Seq models with ChatGPT prompting and showed that few-shot prompting yields superior performance in both DA–MSA and MSA–DA directions. Dialect-specific systems were also explored by (Sibaee et al., 2025), who developed SHAMI-MT for Syrian Arabic using AraT5 and demonstrated high translation quality using LLM-based evaluation. Finally, (Obeidat et al., 2025) introduced IrbidDial, a Jordanian dialect dataset, and showed that few-shot prompting with GPT-based models consistently outperforms fine-tuned NMT systems, even when dialect-aware prompting constraints are applied.

3.2. Influence of POS Tagging and Diacritization on MT

Prior to the widespread adoption of LLMs, several studies investigated the role of linguistic preprocessing in Arabic machine translation. (Baniata et al., 2018) proposed a multitask NMT framework that jointly learns DA–MSA translation and POS tagging, demonstrating that syntactic supervision leads to consistent BLEU improvements for both Levantine and Maghrebi dialects. Complementary work examined the impact of Arabic diacritization on statistical machine translation. (Diab et al., 2007) showed that full diacritization degrades the performance of SMT due to data sparsity, while selective, linguistically motivated schemes produce comparable or slightly improved results. Similarly, (Alqahtani et al., 2016) found that partial diacritization strategies combining lexical and inflectional cues consistently outperform both fully diacritized and undiacritized baselines. Together, these studies indicate that linguistically informed preprocessing can benefit Arabic MT, but overly aggressive annotation increases sparsity and harms performance.

4. Experimental Apparatus

4.1. Procedure

Our experimental procedure is designed to investigate how linguistic features, model capability, and translation metrics affect bidirectional MT between Algerian dialect and MSA. The evaluation framework answers the following research questions:

- We first address **RQ1** in order to measure the effects of POS and diacritics on translation quality and ambiguity resolving.
- We then address **RQ2** by evaluating the ability of LLMs to understand and generate Algerian dialect and MSA.
- Finally, to address **RQ3**, we compare automatic evaluation scores with human judgments across all settings to assess their alignment for dialectal Arabic.

4.2. Dataset

Experiments are conducted on a subset of the PADIC (Meftouh et al., 2015) dataset, which contains sentence-aligned Algerian dialect and MSA pairs and is commonly used for Arabic dialects translation. To incorporate syntactic information, POS tags are added for both language varieties. Algerian dialect POS annotations are generated using the Algerian BERT-based model proposed in (Cheragui et al., 2023), which outputs token-level POS sequences from raw Algerian input. MSA POS tagging is performed using the FARASA toolkit (Darwish and Mubarak, 2016), a state-of-the-art and widely adopted analyzer for MSA. We additionally include diacritics information as an auxiliary linguistic feature. Diacritization for both is performed manually by a linguistic expert to ensure high annotation quality and to avoid noise introduced by automatic diacritization systems.The resulting dataset contains 475 parallel sentence pairs. The average sentence length is 7.02 tokens for Algerian dialect and 8.02 tokens for MSA, revealing a consistent length asymmetry between the two varieties. As shown in introduction, this difference reflects more explicit morphological and syntactic expression in MSA and is a known source of ambiguity in dialect–standard Arabic translation.

4.3. Baseline Models

To evaluate the impact of linguistic features, we selected a set of state-of-the-art LLMs covering proprietary and open-weight models, including GPT-4o-mini, GPT-4o (Latest), and GPT-4-1106-preview from OpenAI (Achiam et al., 2023); Claude-3.5 Sonnet from Anthropic (Anthropic, 2024); Gemini 1.5 Pro from Google (Team et al., 2023); and the open-weight models LLaMA-3.1-8B-Instruct (Grattafiori et al., 2024) and Gemma-2-9B-Instruct (Team et al., 2024). These models span a range of capacities and architectures commonly used in multilingual and translation benchmarks.

This selection enables a controlled comparison of how models with different capacities and training paradigms handle the structural variability and ambiguity characteristic of dialectal Arabic translation. Open-weight models were evaluated locally in quantized configurations using the Ollama inference framework[1] , while proprietary models were accessed through their official APIs: OpenAI[2], Anthropic[3], Google Gemini[4]. All models were assessed using the default hyperparameter settings and the following prompt format.

Translation Prompt Template

Objective: Translate a given sentence from {SOURCE_LANGUAGE} to {TARGET_LANGUAGE}.
Your input will include a sentence in {SOURCE_LANGUAGE} along with: - Diacritization for each word - Part-of-speech tags for each word
Use this information to accurately translate the sentence into {TARGET_LANGUAGE}.
The translation should preserve the original meaning while adhering to {TARGET_LANGUAGE} norms.
Steps:
1. Understand Context: Analyze the sentence, diacritics, and POS tags to grasp its structure and meaning.
2. Word-by-Word Translation: Use diacritization and grammatical class to select appropriate {TARGET_LANGUAGE} equivalents.
3. Reconstruction: Form a grammatically correct and coherent {TARGET_LANGUAGE} sentence.
4. Meaning Verification: Ensure the translation accurately reflects the original intent and nuances.
Output Format: Provide ONLY one translated sentence in {TARGET_LANGUAGE}.

4.4. Metrics

We report results using four automatic evaluation metrics: COMET as a neural-based metric for similarity of vector representations calcualtion,

[1] https://ollama.com/
[2] https://platform.openai.com/
[3] https://platform.claude.com/
[4] https://ai.google.dev/gemini-api

Table 3: Evaluating LLMs on MT with/without POS tags (ALG→MSA). Arrows indicate change relative to W/O. Model suffixes indicate release month.

Model	Comet		MARBERT-v2		BLEU		Chrf++	
	W/O	W/	W/O	W/	W/O	W/	W/O	W/
GPT-4o-mini-07	73.14	69.87↓	75.84	74.04↓	5.44	5.04↓	32.35	30.75↓
GPT-4o-Latest	78.48	77.39↓	79.11	78.34↓	7.95	7.11↓	38.41	36.37↓
GPT-4-1106	72.99	71.59↓	74.6	74.56↓	5.34	5.67↑	30.26	31.86↑
Claude-3.5-sonnet-06	78.57	76.98↓	79.69	79.09↓	8.30	8.56↑	39.28	38.78↓
Claude-3.5-sonnet-10	79.19	77.48↓	80.36	80.03↓	9.02	9.04↑	39.99	39.77↓
Claude-3-sonnet-02	67.12	67.80↑	68.19	69.92↑	3.23	3.55↑	22.80	26.05↑
Claude-opus-02	75.70	74.21↓	77.62	76.77↓	6.49	7.05↑	34.50	33.85↓
Claude-haiku-03	67.59	62.56↓	67.78	64.68↓	3.48	2.64↓	22.82	21.79↓
Gemini 1.5 Pro	76.60	79.19↑	75.63	80.36↑	6.84	9.29↑	34.52	40.34↑
Gemini Flash	69.46	66.73↓	70.76	68.20↓	5.61	4.29↓	29.70	28.68↓
Gemini Flash-8B	70.51	66.91↓	71.07	68.05↓	4.10	2.61↓	27.63	24.51↓
Llama3.1-8B	55.32	50.56↓	62.88	60.06↓	0.63	0.34↓	14.44	13.23↓
Gemma2-9B	64.36	58.63↓	68.18	65.27↓	1.83	1.46↓	20.08	17.60↓

MARBERT-v2 (Abdul-Mageed et al., 2021) (for MSA→ALG) and CamelBERT-MSA (Inoue et al., 2021) (for ALG→MSA) for contextualized semantic similarity, BLEU for n-gram precision, and ChrF++ for character-level overlap. To complement the automated metrics, the human evaluation protocol (Costa-Jussà et al., 2022) looks at fluency and adequacy to ensure the translations sound natural and stay true to the original meaning. Using a simple three-point scale (0, 0.5, 1.0), we evaluate how well models preserve meaning and produce readable translations. Two expert annotators, trained on the evaluation protocol, independently assessed the models' outputs, achieving inter-rater agreement scores of 84.59% for fluency and 86.78% for adequacy.

5. Results and Discussion

5.1. Impact of Linguistic Features (RQ1)

Adding linguistic features was intended to help, but in practice, it often acted as a distraction rather than a helpful signal. Across all experiments, the baseline (w/o features) proved to be the most reliable, where with auxiliary features frequently failed.

- MSA → ALG (Formal to Dialect): in this direction, diacritics (DIACT) offered a signal boost according to human evaluators (Table 7). Adequacy increased for both GPT-4 ($0.88 \rightarrow 0.93$) and Gemini 1.5 pro ($0.91 \rightarrow 0.95$), indicating that vowel markers aid the models in decoding formal Arabic morphology before "translating" it into the dialect. POS tags, on the other hand, were a fluency disaster ($0.93 \rightarrow 0.31$ for GPT-4). Adding more information (POS tokens) causing labels to be mixed into the sentence itself, where the translations of Llama3.1 prove this.

- ALG → MSA (Dialect to Formal): results indicate a strong negative impact in this direction. Performance drops when diacritical marks are introduced into Algerian dialect, a variety that is rarely written with diacritics in real-world usage. For instance, Claude-3.5's BLEU score drops from 9.02 to 3.39. Imposing diacritics on a naturally non-diacritized script shifts the input into a representation unfamiliar to the model, hindering its ability to align dialectal forms with their corresponding MSA version.

- Humans vs. Machines: Humans and LLMs process language differently. Humans use diacritics to reduce ambiguity, while LLMs rely on contextual representations learned from data. Explicit cues such as POS tags can override these learned contextual signals and introduce noise, especially in low-resource settings with limited annotated data.

5.2. Overall Model Performance (RQ2)

A clear distinction in model efficacy is revealed by the comparative analysis. Close-weight models perform better in both configurations, including GPT-4o-Latest and Claude 3.5 Sonnet (20241022). With a COMET score of 79.19 and a ChrF++ of 39.99 in the ALG → MSA direction, Claude 3.5 Sonnet exhibits the highest generalization performance. Significant obvious asymmetry phenomena where models continuously exhibiting higher generation quality when translating from the ALG → MSA direction. Higher BLEU scores for ALG → MSA approximate 9.0, whereas the MSA → ALG direction

Table 4: Evaluating LLMs on MT with/without POS tags (MSA→ALG). Arrows indicate change relative to W/O.

Model	Comet		MARBERT-v2		BLEU		Chrf++	
	W/O	W/	W/O	W/	W/O	W/	W/O	W/
GPT-4o-mini-07	65.88	56.57↓	85.53	86.18↑	2.94	0.18↓	25.82	18.20↓
GPT-4o-Latest	67.67	59.19↓	87.84	88.62↑	5.60	5.52↓	31.16	33.26↑
GPT-4-1106	65.25	58.05↓	84.76	85.17↑	3.30	0.38↓	28.16	20.71↓
Claude-3.5-sonnet-06	68.12	61.85↓	89.49	88.88↓	6.61	3.88↓	33.59	28.45↓
Claude-3.5-sonnet-10	67.56	62.56↓	89.28	90.08↑	7.32	4.05↓	34.11	29.12↓
Claude-3-sonnet-02	59.94	56.49↓	86.64	85.46↓	1.80	0.43↓	22.51	19.24↓
Claude-opus-02	66.55	62.23↓	87.70	87.84↑	5.25	2.33↓	29.69	25.31↓
Claude-haiku-03	62.12	53.78↓	86.65	85.60↓	1.78	0.27↓	20.63	14.95↓
Gemini 1.5 Pro	66.90	60.11↓	87.32	88.64↑	5.25	0.78↓	32.04	22.45↓
Gemini Flash	66.47	58.46↓	87.95	87.68↓	3.58	0.31↓	26.85	19.95↓
Gemini Flash-8B	66.16	57.70↓	87.54	87.11↓	1.73	0.21↓	22.98	18.49↓
Llama3.1-8B	55.13	46.63↓	81.01	76.77↓	0.22	0.05↓	13.86	10.36↓
Gemma2-9B	45.58	50.36↑	79.27	82.68↑	0.13	0.13	4.71	10.77↑

exhibits a marginal decline, peaking at 7.3. Such a disparity suggests that there is a generation bottleneck for low-resource dialects. While the models show sufficient inferential capacity to parse dialectal nuances, they lack the syntactic fluency required for high-quality generation in non-standardized variants.

Additionally, the findings show that smaller open-weights models perform much more poorly. With BLEU scores below 1.0, Llama 3.1 8b and Gemma 2 9b both demonstrate barely useful functionality. Human evaluation metrics (Table 7 and 8) support this functionality collapse, with Llama 3.1 8b reporting adequacy and fluency scores as low as 1.58, demonstrating a near total failure to maintain cross-lingual alignment for this particular pair of languages.

5.3. Metric Analysis (RQ3)

Human assessment of adequacy and fluency (Table 7) indicates an obvious less degradation, whereas automated metrics such as BLEU and ChrF++ demonstrate sharp decreases when diacritical marks are added. A notable discrepancy exists between statistical metrics and human linguistic understanding due to measuring just the overlap with reference without taking semantics into consideration. while most models suffered from performance decay when processing diacritics, Gemini 1.5-Pro and GPT-4 latest demonstrated a positive trend in human-perceived adequacy ($0.91 \rightarrow 0.94$, and $0.88 \rightarrow 0.92$).

MARBERT-v2 scores, which monitored adequacy changes more accurately than traditional MT metrics, notably reflected this beneficial effect as well. The contextual semantic modeling of Arabic in MARBERT-v2 is responsible for this behavior, which makes it especially well-suited for automatically assessing Arabic translation. Overall, these findings show a significant evaluation gap, where improvements in lexical clarity that are easily noticeable by human evaluators are frequently missed by traditional statistical measures.

5.4. Error Analysis

To gain deeper insight into the impact of linguistic augmentation and the quality of LLMs translation, we performed a qualitative error analysis focusing on the output generated by the best performing model (Claude 3.5 sonnet) in each translation direction.

ALG to MSA Without linguistic features most adequacy errors stem from the prevalence of code-switching in Algerian dialect input. In particular, sentences often contain French lexical words (such as باريڤزول, بيان, and لانستيزي) (*for example*, *good*, *anesthesia*) written in Arabic script, which frequently lead to incorrect semantic interpretation. As a result, models tend to either mistranslate these words or ignore their intended meaning, causing substantial semantic drift. Fluency errors are comparatively less frequent in this direction. However, in some cases, outputs retain untranslated Algerian lexical words that often appearing at the beginning of the sentence, indicating incomplete normalization into standard Arabic.

Incorporating linguistic enhancement introduces additional challenges. The presence of dense annotations increases input complexity and introduces noise, which can hinder model comprehension. although outputs typically remain well-formed at the surface level, they often exhibit severe se-

Table 5: Evaluating LLMs on MT with/without Diacritics (MSA→ALG). Arrows indicate change relative to W/O.

Model	Comet		MARBERT-v2		BLEU		ChrF++	
	W/O	W/	W/O	W/	W/O	W/	W/O	W/
GPT-4o-mini-07	65.88	61.36↓	85.53	84.03↓	2.94	2.29↓	25.82	23.84↓
GPT-4o-Latest	67.67	63.58↓	87.84	87.14↓	5.6	4.62↓	31.16	28.68↓
GPT-4-1106	65.25	64.39↓	84.76	88.13↑	3.3	3.88↑	28.16	28.33↑
Claude-3.5-sonnet-06	68.12	64.48↓	89.49	89.22↓	6.61	6.77↑	33.59	31.99↓
Claude-3.5-sonnet-10	67.56	65.07↓	89.28	88.41↓	7.32	7.02↓	34.11	33.41↓
Claude-3-sonnet-02	59.94	57.74↓	86.64	86.36↓	1.8	2.18↑	22.51	20.39↓
Claude-opus-02	66.55	64.57↓	87.7	88.16↑	5.25	5.34↑	29.69	30.1↑
Claude-haiku-03	62.12	58.17↓	86.65	86.53↓	1.78	1.53↓	20.63	18.17↓
Gemini 1.5 Pro	66.9	64.68↓	87.32	88.81↑	5.25	5.52↑	32.04	30.78↓
Gemini Flash	66.47	63.07↓	87.95	88.46↑	3.58	2.74↓	26.85	26.16↓
Gemini Flash-8B	66.16	62.68↓	87.54	88.25↑	1.73	1.2↓	22.98	22.67↓
Llama3.1-8B	55.13	54.79↓	81.01	81.61↑	0.22	0.34↑	13.86	15.78↑
Gemma2-9B	45.58	47.36↑	79.27	84.12↑	0.13	0.19↑	4.71	7.73↑

Table 6: Evaluating LLMs on MT with/without Diacritics (ALG→MSA). Arrows indicate change relative to the W/O setting.

Model	Comet		Camelbert-MSA		BLEU		ChrF++	
	W/O	W/	W/O	W/	W/O	W/	W/O	W/
GPT-4o-mini-07	73.14	48.68↓	75.84	49.65↓	5.44	1.08↓	32.35	9.77↓
GPT-4o-Latest	78.48	54.65↓	79.11	49.78↓	7.95	2.32↓	38.41	15.33↓
GPT-4-1106	72.99	53.61↓	74.6	50.27↓	5.34	1.75↓	30.26	13.9↓
Claude-3.5-sonnet-06	*78.57*	56.23↓	*79.68*	53.24↓	*8.3*	3.59↓	*39.28*	17.83↓
Claude-3.5-sonnet-10	**79.19**	56.08↓	**80.36**	53.42↓	**9.02**	3.39↓	**39.99**	17.45↓
Claude-3-sonnet-02	67.12	43.12↓	68.19	45.59↓	3.23	0.51↓	22.8	5↓
Claude-opus-02	75.7	55.54↓	77.62	52.29↓	6.49	2.96↓	34.5	15.83↓
Claude-haiku-03	67.59	40.68↓	67.78	44.35↓	3.48	0.45↓	22.82	4.19↓
Gemini 1.5 Pro	76.6	54.72↓	75.63	50.67↓	6.84	1.54↓	34.52	13.6↓
Gemini Flash	69.46	53.12↓	70.76	51.18↓	5.61	1.35↓	29.7	11.86↓
Gemini Flash-8B	70.51	49.99↓	71.07	51.42↓	4.1	0.78↓	27.63	6.33↓
Llama3.1-8B	55.32	44.37↓	62.88	46.72↓	0.63	0.06↓	14.44	5.39↓
Gemma2-9B	64.36	47.94↓	68.18	50.22↓	1.83	0.51↓	20.08	8.18↓

mantic mismatches, producing grammatically correct sentences that are unrelated to the original input content or topic.

MSA to ALG Without linguistic features, several recurrent error types are observed. First, outputs sometimes retain MSA vocabulary instead of generating dialectal equivalents, reflecting incomplete dialect adaptation. Second, models occasionally produce lexical words from other Arabic dialects, most commonly Moroccan dialect.

Orthographic distortions are also frequent, where minor character alterations lead to incorrect word forms (e.g., generating تعام (*food*) instead of طعام or خطارت (*choose*) instead of اخترت). Additionally, models sometimes reorder internal characters or syllables, resulting in awkward or nonstandard pronunciation patterns.

Another common issue involves grammatical agreement, particularly gender mismatches, where model incorrectly switch between masculine and feminine forms during translation.

When POS tags are introduced, a distinct class of semantic errors emerges. Models sometimes confuse syntactic roles, especially between subjects and objects, leading to outputs that are linguistically fluent and dialectically appropriate but semantically incorrect. This indicates that while linguistic features may improve surface quality, they can also disrupt deeper sentence-level interpretation.

6. Conclusion

This study used LLMs to investigate how linguistic augmentation affects bidirectional translation between Algerian dialect and MSA. Findings indicate that auxiliary features like POS tags and diacritics frequently cause noise instead of enhancing performance, with asymmetric effects in both directions:

Table 7: Human evaluation for MSA→ALG (Adequacy and Fluency). Arrows indicate change relative to w/o.

Model	Adequacy			Fluency		
	w/o	w/ POS	w/ DIACT	w/o	w/ POS	w/ DIACT
GPT-4 models*	88.63	34.74↓	92.95↑	93.05	31.26↓	87.26↓
Claude-3-5-sonnet-10	93.68	67.05↓	88.74↓	88.74	58.84↓	81.37↓
Google-gemini-1.5-pro	91.46	45.26↓	94.53↑	87.58	37.05↓	87.26↓
Gemma2:9b-instruct-q5	4.74	3.37↓	4.00↓	4.74	2.32↓	2.21↓
Llama3.1:8b-instruct-q5	1.58	1.47↓	2.84↑	1.58	0.42↓	3.05↑

Table 8: Human evaluation for ALG→MSA (Adequacy and Fluency). Arrows indicate change relative to w/o.

Model	Adequacy			Fluency		
	w/o	**w/ POS**	**w/ DIACT**	**w/o**	**w/ POS**	**w/ DIACT**
GPT-4 models*	82.70	65.05↓	45.26↓	95.05	56.42↓	38.00↓
Claude-3-5-sonnet-10	80.42	67.58↓	72.53↓	96.42	59.26↓	65.89↓
Google-gemini-1.5-pro	52.32	61.68↑	53.89↑	90.74	57.68↓	43.37↓
Gemma2:9b-instruct-q5	6.84	6.00↓	5.26↓	6.95	4.42↓	4.42↓
Llama3.1:8b-instruct-q5	1.89	0.21↓	0.53↓	1.89	0↓	0.32↓

diacritics can improve adequacy in MSA→ALG but significantly hinder translation in ALG→MSA because of distribution mismatch. Additionally, we see distinct differences in capability between models: smaller open-weight models exhibit significant performance degradation, while frontier LLMs show strong dialect comprehension but are still limited in dialect generation. Lastly, a significant discrepancy between automated metrics and human assessment is revealed, with contextual semantic metrics outperforming traditional overlap-based measures in capturing adequacy changes.

7. References

Ahmed Elmogtaba Abdelmoniem Ali Abdelaziz, Ashraf Hatim Elneima, and Kareem Darwish. 2024. Llm-based mt data creation: Dialectal to msa translation shared task. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation@ LREC-COLING 2024*, pages 112–116.

Muhammad Abdul-Mageed and Houda Bouamor. 2021. Arabic text processing with morphological and diacritical features. *Computational Linguistics*.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, et al. 2021. Arbert & marbert: Deep bidirectional transformers for arabic. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 7088–7105.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Abdullah Alabdullah, Lifeng Han, and Chenghua Lin. 2025. Advancing dialectal arabic to modern standard arabic machine translation. *arXiv preprint arXiv:2507.20301*.

Sawsan Alqahtani, Mahmoud Ghoneim, and Mona Diab. 2016. Investigating the impact of various partial diacritization schemes on arabic-english statistical machine translation. In *Conferences of the Association for Machine Translation in the Americas: MT Researchers' Track*, pages 191–204.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1.

Hanin Atwany, Nour Rabih, Ibrahim Mohammed, Abdul Waheed, and Bhiksha Raj. 2024. Osact 2024 task 2: Arabic dialect to msa translation. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation@ LREC-COLING 2024*, pages 98–103.

Baligh Babaali, Mohammed Salem, and Nawaf R Alharbe. 2025. Breaking language barriers with

chatgpt: enhancing low-resource machine translation between algerian arabic and msa. *International Journal of Information Technology*, 17(7):4109–4118.

Laith H Baniata, Seyoung Park, and Seong-Bae Park. 2018. A multitask-based neural machine translation model with part-of-speech tags integration for arabic dialects. *Applied Sciences*, 8(12):2502.

Reem Bassiouney. 2009. *Arabic Sociolinguistics*. Edinburgh University Press.

Mohamed Benrabah. 2014. Competition between four "world" languages in algeria. *Journal of World Languages*, 1(1):38–59.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, et al. 2018. The madar arabic dialect corpus and lexicon. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Farida Bouhadiba. 2017. Algerian arabic: A sociolinguistic overview. In Elabbas Benmamoun and Reem Bassiouney, editors, *The Routledge Handbook of Arabic Linguistics*, pages 353–368. Routledge.

Mohamed Amine Cheragui, Abdelhalim Hafedh Dahou, and Amin Abdedaiem. 2023. Exploring bert models for part-of-speech tagging in the algerian dialect: A comprehensive study. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 140–150.

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

Anna Currey and Kenneth Heafield. 2020. Incorporating linguistic annotations into neural machine translation. In *Proceedings of WMT*.

Kareem Darwish and Hamdy Mubarak. 2016. Farasa: A new fast and accurate arabic word segmenter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1070–1074.

Mona Diab, Mahmoud Ghoneim, and Nizar Habash. 2007. Arabic diacritization in the context of statistical machine translation. In *Proceedings of Machine Translation Summit XI: Papers*.

Mona Diab and Kadri Hacioglu. 2004. Automatic diacritization of arabic text. In *Proceedings of COLING*.

Lukas Edman and Anders Sogaard. 2023. Do large language models benefit from linguistic supervision? In *Proceedings of ACL*.

Ashraf Hatim Elneima, AhmedElmogtaba Abdelmoniem Ali Abdelaziz, and Kareem Darwish. 2024. Osact6 dialect to msa translation shared task overview. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation@ LREC-COLING 2024*, pages 93–97.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Nizar Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool.

Nizar Habash. 2013. *Arabic Natural Language Processing*. Synthesis Lectures on Human Language Technologies.

Nizar Habash. 2019. *Arabic Natural Language Processing*. Cambridge University Press.

Nizar Habash and Owen Rambow. 2007. Arabic diacritization through full morphological tagging. In *Proceedings of NAACL*.

Clive Holes. 2004. *Modern Arabic: Structures, Functions, and Varieties*, revised edition. Georgetown University Press.

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*.

Salam Khalifa et al. 2022. Camel tools: An open platform for arabic natural language processing. In *Proceedings of LREC*.

Abdullah Salem Khered, Youcef Benkhedda, and Riza Theresa Batista-Navarro. 2025. Dial2msa-verified: A multi-dialect arabic social media dataset for neural machine translation to modern standard arabic. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 50–62.

Djamel Eddine Kouloughli. 1999. *Le maghrébin*. Presses Universitaires de France.

Yinhan Liu and Graham Neubig. 2024. Revisiting explicit syntax in neural machine translation. *Transactions of the ACL*.

Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas, and Kamel Smaili. 2015. Machine translation experiments on padic: A parallel arabic dialect corpus. In *Proceedings of the 29th Pacific Asia conference on language, information and computation*, pages 26–34.

Rasha Obeidat, Luay Alawneh, and Yara Al-Harahsheh. 2025. Exploring prompting for dialectical machine translation: a focus on north jordanian arabic. *PeerJ Computer Science*, 11:e3209.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of LREC*.

Lotfi Sayahi. 2014. *Diglossia and Language Contact: Language Variation and Change in North Africa*. Cambridge University Press.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91.

Serry Sibaee, Omer Nacar, Yasser Al-Habashi, Adel Ammar, and Wadii Boulila. 2025. Shami-mt: a syrian arabic dialect to modern standard arabic bidirectional machine translation system. *arXiv preprint arXiv:2508.02268*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Kees Versteegh. 2014. *The Arabic Language*, 2 edition. Edinburgh University Press.

Nasser Zalmout and Nizar Habash. 2020. Adversarial neural arabic diacritization. In *Proceedings of ACL*.

Imed Zitouni, Jeffrey Sorensen, and Ruhi Sarikaya. 2006. Arabic diacritization using a statistical approach. In *Proceedings of HLT-NAACL*.