HurtNet: a Multilingual Dictionary of Hurtful Words

Anonymous ACL submission

Abstract

We study the problem of preserving offensive load in LLM-generated translations, focusing on how offensive words retain their load across languages. We draw on HurtNet, a novel language resource consisting of five multilingual (Arabic, French, Bulgarian, Greek, and Italian) dictionaries of hurtful terms, with multiple definitions and examples of usage. We draw on HurtNet to identify offensive senses shared by the 5 languages and the terms that denote them, and feed translations with this knowledge. We evaluate the approach through quantitative and qualitative analysis. A series of knowledge-injection zero-shot translation experiments were carried out with four different prompts that gradually increase injected linguistic knowledge, and three LLMs – an open use, a translation-dedicated, and a commercial one. Our results show that translations better preserving offensive load are produced when the prompt explicitly instructs the model about the translated offensive term and that the method is beneficial for those cases where the literal translation of the term does not preserve its offensive load.

Keywords: Machine Translation, Offensive Language, Multilingual Dictionary of Offensive Language

1 Introduction

Machine translation (MT) is a primary tool for making social media content accessible across language barriers. As platforms grow increasingly multilingual, MT systems are routinely deployed to translate user-generated content at scale — exposing millions of users to automatically translated posts, comments, and interactions. However, this content is inherently challenging to translate: among the many reasons that lead to mistranslations, idiomatic language, and cultural differences between source and target language play an important role.

These limitations become particularly severe when dealing with offensive and hurtful language[1]. MT models frequently produce outputs that neutralize, distort, or amplify the offensive load of the source expression with serious consequences for downstream tasks such as sentiment analysis, content moderation, and offensive language detection. Similar challenges have been documented in cross-lingual hate speech classification, where the transfer of offensive language detectors between languages suffers from comparable distortions (Dmonte et al., 2025). The impact of MT quality on these tasks remains largely under-explored.

Social media platforms have greatly facilitated the dissemination of hateful and offensive speech and have prioritized the need to mitigate the spread of online hate and, consequently, to automatically identify offensive material. To this end, a growing body of research has focused on developing linguistic resources for automatic offensive language detection (Poletto et al., 2020).

Since the core of hurtful speech ultimately resides in lexical choice, alongside annotated corpora, lexicons are used to capture lexical indicators of offensiveness. HurtLex (Bassignana et al., 2018) stands out as the first multilingual offensive lexicon. It was designed to provide a structured and language-independent inventory of hurtful words in multiple languages, thus facilitating cross-lingual research in hate speech detection (Pamungkas and Patti, 2019; Monnar et al., 2024).

Nevertheless, listing a term as offensive presents the problem of polysemy. It is often the case that only some of the multiple senses of a polysemous word carry derogatory or offensive meaning (e.g., the word "heat" has both the meaning of "the quality of being hot or warm, or the temperature of

[1] We use the terms *hurtful* and *offensive* interchangeably to refer to language that may cause harm or be perceived as disrespectful, since both terms encompass a spectrum of hurtful language phenomena (Poletto et al., 2020).

something" and "a state of sexual excitement and readiness to breed" used for animals or, in a derogatory sense, for people). When an MT system fails to identify the offensive sense of an ambiguous term, the resulting translation may lose, alter, or inadvertently introduce offensive content. In the same vein, hate speech detection models must distinguish between neutral and hateful usages of the same ambiguous lexical item (Muti et al., 2024; Muzakir et al., 2023; Taylor et al., 2017). Therefore, a resource providing explicit definitions of the derogatory senses of each entry becomes essential to ensure more accurate semantic interpretation and contextual disambiguation.

In this work, we address this gap by relying on HurtNet, which is a novel dictionary built on top of HurtLex. HurtNet enriches its entries with detailed sense definitions and contextualized usage examples in five languages. HurtNet is the first multilingual, semantically informed dictionary of offensive language of its kind, to the best of our knowledge. We use HurtNet to identify derogatory senses shared by all five languages (see Section 3) and, drawing on them, evaluate the resource through a machine translation framework, treating MT as both a stress test and a practical use case: if a semantically grounded offensive dictionary can improve the identification and handling of derogatory language in translation, it holds significant promise for cross-lingual offensive language processing more broadly.

2 Related Works

It has already been pointed out that offensive language is a significant challenge for machine translation systems. Offensive expressions are very often idiomatic, and their emotional weight varies considerably between cultures. Models tend to translate idioms literally, causing the intended offense to be lost or softened (Ávila-Cabrera, 2023; Zhao et al., 2026). In addition, MT systems are prone to "added toxicity", a phenomenon where the model introduces toxic or offensive terms into the generated output, even when the source text is neutral (Costa-jussà et al., 2023). Lastly, there is a frequent amplification of existing societal biases through the generation of stereotypical translations that disproportionately affect marginalized groups, especially when translating low-resource languages (Haroutunian, 2022; Yao et al., 2025).

Lexicon-based approaches for the detection of hateful content offer notable advantages due to their transparency, traceability and light computational burden: researchers and practitioners can easily inspect which words or expressions contribute to classification decisions, adapt or expand the lists to new domains or languages, and maintain interpretability, an aspect often lacking in black-box automated approaches. Albadi et al. (2018) automatically generated three Arabic lexica using three feature selection methods. Also for Arabic, Mubarak et al. (2017) constructed a lexicon to support the detection of hate speech in Arabic social media. For English, Davidson et al. (2017) compiled a list of 179 hate-related words derived from the user-generated HateBase [2] database is an online repository of multilingual, usage-based hate speech with 3894 terms in 98 languages. Wiegand et al. (2018) introduced two lexica of abusive terms: a base lexicon and an expanded one with $1,650$ and $8,478$ entries respectively, where each word is annotated as abusive or non-abusive. Finally, PeaceTech Lab (Ferroggiaro, 2018) released a multilingual hate speech lexicon as part of a humanitarian initiative in Central Africa, including hate-related terms and their alternative spellings in several languages (English, Fulani, Hausa, Igbo, Pidgin, and Yoruba).

2.1 HurtLex

HurtLex (Bassignana et al., 2018) is a multilingual computational lexicon of hate terms that started from the Italian lexicon "Le Parole per Ferire" by (De Mauro, 2016b), and automatically translated into more than 50 languages. The starting lexicon comprises more than $1,000$ Italian words organized into three main macrocategories subdivided into seventeen semantic categories: derogatory terms, words bearing stereotypes, and neutral words that can acquire an offensive meaning through semantic shift.

As a freely available language resource, HurtLex has been the object of improvements and refinements. Tontodimamma et al. (2023) augmented Italian HurtLex with offense scores computed with Item Response Theory. Osenova (2024) reports on a systematically cleaned and normalized Bulgarian HurtLex. Stamou et al. (2022) improved Greek HurtLex with manual annotation on context, reference, and thematic domain, and Stanković et al. (2020) refined Serbian HurtLex with special focus

[2] https://hatebase.org/

on the integration of multiword expressions.

HurtLex was used as a resource for offensive language detection in several research works: as a straightforward look-up component for offensive terms (Giordano and Di Buono, 2023), as a method to extract task-relevant lexical features (Árcos and Pérez, 2023), and as a component of more sophisticated neural architectures (Koufakou et al., 2020).

3 HurtNet: from Multilingual Lexicons to a Knowledge Resource

Each HurtNet component was built from the corresponding language version of HurtLex, either the core version or an already improved one, as in the case of Greek. This was followed by a language-specific validation process, which produced notable variations in entry counts across languages (Table 1): some entries were removed as non-offensive, others merged under a single entry, and in certain cases, most notably Greek, new entries were added upon discovery of additional offensive senses.

HurtNet integrates these independently developed offensive-language dictionaries into a single, coherent multilingual resource freely available to the public. The starting point for this integration was the definition of a shared data model that all five language components conform to.

Three main information fields structure each entry in the unified dataset. The first contains the core information shared across all languages: a unique identifier (the HurtNet ID), the lemma, a definition of its offensive use, and a usage example, each with their respective sources. The second field links HurtNet to external lexical resources: where available, each lemma is associated with a synset from the Open Multilingual Wordnet (OMW), and the corresponding synset ID and synset lemma are recorded. The third field accommodates language-specific extensions. For instance, the Greek component includes an offensiveness score derived from AIKIA (Markantonatou et al., 2024) and ERIS (Basile et al., 2023) data; the Arabic component provides a reference to the stereotype associated with each entry; and the Bulgarian component specifies the synset category. The harmonized, multilingual version of the resource, comprising 2786 entries, is available on Zenodo.[3]

[3] https://zenodo.org/records/17414084

Table 1: Statistics of HurtNet: number of original entries, final entries, retention rate, and the final number of entries.

Language	Original entries	Final entries	Retention (%)
Arabic	501	437	87.22%
Bulgarian	1370	637	46.50%
French	501	331	66.07%
Greek	737	1109	150.47%
Italian	543	272	50.09%
Hurtnet		**2786**	

3.1 Shared semantic categories in HurtNet

HurtNet entries were grouped into global senses using an LLM-as-a-judge approach for interlingual alignment. Since comparisons by the model lacked inherent transitivity, global senses were extracted via a strict pairwise merging rule, by which a 5-language group was accepted only if all 10 internal pairs merged. These groups were then consolidated into 31 distinct categories by a 4/5 rule: if four of the five languages merged, the differing fifth language was treated as an alternative lexical choice rather than a separate meaning.

3.2 Arabic

The development of the HurtNet Arabic component was carried out by a native speaker of Levantine Arabic. Entries in the Arabic HurtLex core, which contained 501 entries in Modern Standard Arabic (MSA), were examined to determine whether they conveyed a potentially offensive sense, in both MSA and Levantine Arabic varieties. Entries that did not meet this criterion and duplicate entries were eliminated, resulting in a reduced lexicon of 437 entries. After the correction of spelling errors and incorrect part of speech labels, the lexicon was lemmatized, with special attention given to word forms whose inflection produced a semantic shift affecting offensiveness, for instance, the plural رُعَاع *ru'ā'* "disreputable people" or the feminine most common form شَرموطة *sharmuta* "whore". Arabic WordNet (AWN v2) was then used to map the entries to their corresponding synsets. In total, 193 items were successfully aligned, and their synset identifiers were integrated into the lexicon, while the remaining entries for which no suitable synset was found were left without an identifier. Given the scarcity of publicly available Arabic lexical resources, LLMs were used to generate concise offensive definitions and illustrative examples written in simple MSA to ensure broad intelligibility.

All generated content was manually reviewed and corrected to ensure the accuracy and relevance of the offensive usage.

3.3 Bulgarian

Version 1.2 for Bulgarian HurtLex, containing 2865 terms, was downloaded from the multilingual HurtLex repository and manually checked. After deletion of repetitions and the normalization of identical terms with varying spelling —mostly the same term with a capital and small letter — the lexicon consists of 1370 terms (slightly more than 50% of the original list of terms).

The lexicon was further enriched with synonyms in the appropriate synsets through sense mappings with BTB-Wordnet (Simov and Osenova, 2023). Around 600 lemmas have direct mappings to BTB-Wordnet, while 370 lemmas are linked to the English Wordnet 3.0.

The technically cleaned Bulgarian Hurtlex was inserted into a customized version of the specially designed CLaDA-BG Dictionary Creation System (Angelov et al., 2022). This version includes the following information: the lexeme/phrase, the part-of-speech of the lexeme or of the headword of the phrase, the status of the entry completion (with labels Ready, To check, Irrelevant), indication of whether the term is present in WordNet or not (with labels True or False), the definition of the lexeme meaning (it comes from the WordNet —if it is present there, or from other available dictionaries — if the lexeme and/or its appropriate meaning is absent), and comments on any issues related to the check-up. Finally, the main sources of the examples are the BTB-WordNet and Bulgarian CLASSLA web corpus[4].

3.4 French

The French version of HurtLex core originally contained 501 entries. Their filtering, annotation, and validation were carried out by two master's students in linguistics under academic supervision. Each lexical item underwent manual review to verify whether the entry reflected one of three types of usage: offensive (explicitly hateful or injurious), non-literally pejorative (pejorative by semantic deviation, such as metaphorical use), or negative connotation (bearing a negative evaluative meaning without direct insult). Entries that did not meet these criteria were excluded, resulting in a refined subset of 331 words. Inflected forms were lemmatized, and only variants with genuine semantic shifts were retained—for example, *pouffiasse* (Eng. *slut*), a colloquial derivative of *pouf* (Eng. *tart*), used to demean a woman perceived as vulgar or provocative.

French HurtNet lemmas were mapped on the corresponding synsets in Open Multilingual WordNet (OMW), Version 1.4. In total, 248 items were aligned, and their associated information —the OMW identifier and lemmas— was integrated into the dictionary. Entries without a corresponding match were left with empty OMW identifier and lemma fields.

The `definition` column was populated with entries from the OMW. When no definition was available or when the existing one was deemed contextually inadequate, complementary definitions were retrieved from Le Robert[5], a leading reference dictionary of the French language. All collected definitions were systematically reviewed to ensure semantic adequacy and precision, and to eliminate definitions reusing the lexical item being defined. The provenance of each definition was documented in the `definition' source` column.

In the `example` column each lemma was associated with an authentic or representative usage instance. When available and contextually appropriate, examples were drawn from OMW. Otherwise, corpus-based data from the Sketch Engine platform[6] were used, prioritizing examples that reflect real-world usage and intent. If corpus evidence was missing or unsatisfactory, examples were created manually or generated using LLMs (ChatGPT 4o). The source of each example was systematically indicated in the `example' source` column.

3.5 Greek

Modern Greek (MG) currently lacks a standardized dictionary of offensive vocabulary. To fill this gap, we built a sense-level MG offensive dictionary from the 737 terms of the cleaned Greek HurtLex (Stamou et al., 2022). GPT-o3 was prompted to generate sense-separated genus–differentia definitions (Atkins and Rundell, 2008), three GDEX-style usage examples per sense (Kilgarriff et al., 2008), and a target class from a six-way CIDOC-CRM-inspired controlled vocabulary (Person, Object, State, Event/Result, Repeated Action, Behavioral Trait) (CIDOC CRM Special Interest Group, 2021).

[4] https://www.clarin.si/ske/#dashboard?corpname=classlaweb2_bg

[5] https://www.lerobert.com/

[6] https://www.sketchengine.eu/

The 802 LLM drafts were post-edited (536 logged edits), yielding 1,109 final sense entries. Examples were drawn from the AIKIA corpus of MG offensive speech (Markantonatou et al., 2024); the ∼200 senses not attested in AIKIA were illustrated with Twitter/X instances scored via Best–Worst Scaling (Basile and Cagnazzo, 2021), matching the AIKIA protocol. Each entry carries an offense score adopted from the ERIS lexicon (Basile et al., 2023). An evaluation of 20 entries (32 senses) by 3 lexicographers yielded an overall score of 81%, with recurring weaknesses such as archaic forms presented as current, English-pattern interference, evaluative-tone leakage, and sense drift in generated examples.

3.6 Italian

The dictionary construction started from the 543 offensive terms provided by Hurtlex core. Each term was manually reviewed to verify the presence of offensive or derogatory content within its definitional or usage context. Terms for which only neutral definitions could be identified, and for which no offensive evidence was found, were excluded. After filtering, 272 items were retained.

To integrate the entries with the OMW, all HurtLex terms were matched to OMW synsets whenever possible. For each term, the corresponding `synset_id` and `synset_lemmas` were recorded. When no synset was available but both an offensive definition and an example were found, the entry was included with placeholder values (`[no_synset]`, `[no_synset_lemma]`). Duplicate entries, typically corresponding to gender or number variants, were removed.

Definitions were first retrieved from OMW and stored in the `definition` column. When these definitions were either non-offensive or insufficient to capture the offensive sense of the term, alternative definitions were drawn from the Nuovo De Mauro (De Mauro, 2016a) and Sabatini-Coletti (Sabatini and Coletti, 2018) dictionaries, or, when necessary, manually formulated according to lexicographic principles. All definitions underwent manual validation, and their origin was documented in the `'definition' source` column.

Since OMW does not provide Italian examples, the `example` column was completed using data extracted primarily from SketchEngine (in the resource *se*), supplemented by entries from Nuovo De Mauro (*dm*) or manually created examples when corpus evidence was lacking (*hn*). The `example' source` column systematically records the source of each instance.

4 Experimental Validation

A machine-translation experiment was designed to assess whether the lexical and alignment information provided by HurtNet helps large language models preserve a term's offense load (sense, target, severity, and pragmatic function) when translating offensive expressions across HurtNet languages.

4.1 Gold set

In order to develop a gold set of 200 sentences against which translation results would be evaluated, 40 terms were selected representing the 31 global-sense categories of Section 3.1, and 40 sentences from the dictionary's usage examples. Each example featured a single offensive term and contextual cues that disambiguated its intended sense (Isabelle et al., 2017); this was a necessary choice because offensive terms require sense-level rather than surface-level translation (e.g., the Greek offensive use of *donkey* for moral corruption maps to English *jerk*, not *donkey*, whose offensive sense is 'stupid'). The 31 global-sense categories represent semantic domains such as low intelligence, deception, maliciousness, worthlessness, sexualized insults, ideological insults, animal-based insults, and chaos/dirtiness, and vary in part of speech, target type, and degree of offensiveness. Each sentence was rendered in all five HurtNet languages (Arabic, Bulgarian, Modern Greek, French, Italian) by a native speaker and reviewed by a second native speaker per language, yielding 200 parallel sentences. Each entry carries metadata in line with Data Statements (Bender and Friedman, 2018): source sentence, target-language translations, offensive term, sense identifier, offensive cluster, target type, and POS.

4.2 Methodology

We selected three large-language models: gemma-3-12b-it (Team, 2025) (open source), MiLMMT-46-12B-v0.1 (Shang et al., 2026) (open source multilingual MT model) and GPT-5.4-mini (commercial) [7].

The MT experiment comprised four prompting conditions, each representing a different level of

[7] For the open-source models, inference was run with vLLM using $temperature = 0$, $top_p = 1.0$, and $max_tokens = 300$.

information supplied to the model. In each condition, the model was asked to return JSON with two fields: The `Sentence_Target` field containing the translated sentence in the target language and the `Reason` field, containing a short English explanation of the translation choices.

The instructions also specified constraints against formatting and content inconsistencies in the generated outputs (See Appendix A).

In the **Baseline** condition, a plain translation instruction is given to the model without explicitly mentioning that the sentence contains offensive language.

In the **Offense-Aware** condition, the prompt explicitly informed the model that the sentence contained an offensive expression and asked it to preserve its offensive load during translation.

In the **Sense-Informed** condition, the prompt supplied the original offensive term, its definition in the source language, and verified aligned candidates in the target language 3. The model was instructed to select the candidate that most closely matched the offense load of the original offensive expression within the given context.

In the **Source+Target Lexicon** condition, the model received the full offensive HurtNet dictionary for both the source language and the target language. Unlike the **Sense-Informed** condition, the model was not given the source term, source sense ID, or source definition as separate metadata. Instead, it had to identify the offensive expression and its contextual sense by searching the source-language HurtNet dictionary, and then select the closest corresponding sense from the target-language HurtNet dictionary. This condition was designed to test whether the model can use HurtNet as an external lexical resource for both source-side sense identification and target-side lexical choice.

After collecting the output of each experiment, systematic errors were identified and cleaned manually or automatically. The errors fall into four main groups and mainly concern the last condition: empty fields (885), labels included in the translated sentence (747), explanations placed in the wrong field (1741), and wrong-language or mixed-language outputs (139).

4.3 Results

For evaluation, the models translated each gold sentence (Section 4.1) from one source HurtNet language into the other four languages under each prompt condition. The outputs were then scored at the sentence level against the corresponding gold reference using four complementary metrics. BLEU and chrF++ (computed via sacrebleu[8]) measure lexical overlap with the reference translation through n-gram precision and character-level matching, respectively, with chrF++ additionally accounting for word order. Semantic similarity is captured by mBERT F1 (bert-score with bert-base-multilingual-cased), while overall neural MT quality is estimated by COMET (wmt22-comet-da), both operating on a 0–1 scale.

Table 2: Average translation quality scores across prompting strategies and models. Best result for each experiment is shown in bold.

Experiment	Model	n	BLEU	chrF++	mBERT F1	COMET
Baseline	Gemma-3 12B	800	27.63	53.01	0.8392	0.8630
Baseline	MiLMMT-46 12B	800	28.66	55.90	0.8450	0.8742
Baseline	GPT-5.4-mini	800	**29.55**	**56.58**	**0.8498**	**0.8779**
Offense-Aware	Gemma-3 12B	800	23.39	50.37	0.8244	0.8264
Offense-Aware	MiLMMT-46 12B	800	28.33	55.61	0.8443	0.8724
Offense-Aware	GPT-5.4-mini	800	**29.73**	**56.72**	**0.8485**	**0.8751**
Sense-Informed	Gemma-3 12B	800	27.14	53.00	0.8408	0.8572
Sense-Informed	MiLMMT-46 12B	800	29.00	55.79	0.8458	0.8750
Sense-Informed	GPT-5.4-mini	800	**31.13**	**56.95**	**0.8568**	**0.8764**
Source+Target Lexicon	Gemma-3 12B	800	25.61	51.37	0.8338	0.8598
Source+Target Lexicon	MiLMMT-46 12B	800	29.12	55.71	0.8478	0.8756
Source+Target Lexicon	GPT-5.4-mini	800	**30.10**	**56.39**	**0.8512**	**0.8781**

GPT-5.4-mini returned the best MT results overall (see Table 2), with the **Sense-Informed** condition giving the best scores on BLEU, chrF++, and mBERT F1. COMET did not always follow the same trend, as it is sensitive to broader fluency and adequacy rather than offensive-word choice alone. Bootstrap confidence intervals confirm that **Sense-Informed** prompting yielded the most consistent BLEU gains for GPT-5.4-mini overall ($\Delta = +1.58$, 95% BCa CI $[+0.55, +2.72]$). As visible in Figure 1, the most pronounced improvements are concentrated in language pairs targeting Greek. The **Source+Target Lexicon** condition yielded moderate but generally positive gains. In contrast, the **Offense-Aware** condition was the least stable: it frequently degraded performance — particularly when translating into Arabic or French — suggesting that merely flagging offensive content without structured lexical guidance can interfere with translation quality rather than improve it[9].

At language level, Arabic returned the lowest BLEU, chrF++, and COMET scores. Bulgarian returned the highest COMET, while Italian and Bulgarian showed the highest surface-overlap scores.

[8] https://github.com/mjpost/sacrebleu

[9] Further analysis on other metrics and models can be found in Appendices B and D.

Figure 1: Per-language-pair delta scores relative to the **Baseline** for GPT-5.4-mini

Greek returned lower BLEU but high COMET, which suggests that several Greek outputs may be semantically acceptable even when they differ lexically from the reference.

As source languages, French and Italian returned the highest COMET averages and Greek the lowest one. This suggests that some Greek source cases may be harder to transfer to the HurtNet target languages, especially when the offensive term is idiomatic or metaphorical[10].

5 Qualitative Analysis

In addition to automatic metrics, a qualitative analysis was performed. Three categories of offensive terms were identified among the 40 selected to develop the gold set (Section 4.1):

1. Common Offensive Words - Literal Translation: offensive senses shared by the two languages, and the offensive terms are literal translations of each other, for example, 'fascist' or 'lie'.

2. Different Offensive Word - Same Offense Load: a non-literal translation of the target word carries the required offense load; this includes synonym-like cases.

3. Changed Offensive Word - Changed Offense Load: the translated word changes the offense load of the offensive expression.

With **Sense-Informed** and **Source+Target Lexicon** prompts, across all three categories, roughly 70% of the sentences are either improved or remain neutral compared to **Baseline**. The gains are mainly discussed with reference to the COMET metric, which provides a neural estimate of translation quality in the $0-1$ range. This improvement is especially relevant for the last group, where we expected the richer prompts to be more effective.

Richer prompts yielded limited gains for common insults (e.g., 'lie' or 'fascism'), but can help substantially with sexual expressions, false friends, metaphorical insults, and highly specific slang.

The example in Table 3 shows a classic false-friend case: French *baiser* "to fuck" in this context was mistranslated in the Italian **Baseline** as *baciare* "to kiss". Both **Sense-Informed** and **Source+Target Lexicon** corrected the error by selecting *scopare* "to fuck" (COMET gain: $+0.206$).

Table 3: Example of a false-friend translation case across prompting strategies in Gemma 3 12B.

Source (FR)	*tu as baisé la brésilienne?*	–
Reference (IT)	*Hai scopato la brasiliana?*	–
Reference (EN)	*Did you fuck the Brazilian woman?*	–
Prompt	**Translation**	**COMET**
Baseline	*Hai baciato la brasiliana?* (“Have you kissed the Brazilian woman?”)	0.725
Sense-Informed	*Hai scopato con la brasiliana?* (“Have you fucked the Brazilian woman?”)	**0.930**
Lexicon	*Hai scopato con la brasiliana?* (“Have you fucked the Brazilian woman?”)	**0.930**

Table 4: Example where the translations remove the offensive expression entirely across prompting strategies in Gemma 3 12B.

Source (EL)	μεγαλύτερη κουταμάρα από τη μεταβίβαση του ποσοστού δεν μπορούσε να σκεφτεί ο σεναριογράφος	–
Reference (FR)	*Le scénariste n'aurait pas pu imaginer une plus grande connerie que le transfert de l'action*	–
Reference (EN)	*The screenwriter couldn't have come up with a bigger absurdity than the transfer of the share*	–
Prompt	**Translation**	**COMET**
Baseline	*Le scénariste ne pouvait imaginer pire que le transfert du pourcentage.* (The screenwriter couldn't imagine anything worse than the transfer of the percentage.)	0.672
Sense-Informed	*Le scénariste n'a pas pu imaginer plus grande connerie que le transfert du pourcentage.* (The screenwriter couldn't have imagined a bigger bullshit than the transfer of the percentage.)	0.847
Source+Target Lexicon	*Le scénariste ne pouvait imaginer plus grande bêtise que de céder le pourcentage.* (The screenwriter couldn't imagine a greater stupidity than giving up the percentage.)	**0.873**

In the example in Table 4, the offensive term does not occur in the **Baseline**. **Sense-Informed** used the definition of Greek κουταμάρα *koutamára*

[10] See Table 5 and 6 in Appendix C.

"stupidity/foolishness" to select French *connerie* "bullshit/stupidity", which better preserves vulgar nuance. **Source+Target Lexicon** selected *bêtise* "foolishnes", which is semantically close but less vulgar (COMET gain: $+0.175$).

The examples given above illustrate the observed pattern where richer prompts help most when the offensive expression is ambiguous, metaphorical, culturally loaded, or easily confused with a non-offensive meaning. So, entries of type (2) and (3) consistently benefit MT performance for offensive terms, entries of type (1) provide additional information that proves superfluous for the task, and entries of type (4) call for closer inspection:

1. Common general insults: about half of these cases are unaffected because the models already know terms such as "idiot" or "stupid"; the remaining gains and losses are of similar magnitude.

2. Sexual expressions and false friends: **Sense-Informed** gives large gains, often around $+0.20$ to $+0.28$ COMET, because the model receives sense-level evidence instead of relying only on surface form.

3. Metaphorical uses: expressions such as μπουρδέλο *bourdelo* as "chaos", "ordure" as "despicable person", or σκούπιδι *skoupidi* "trash" as a human-directed insult benefit strongly from **Sense-Informed** prompting.

4. Very strong insults with specific targets, including racist or homophobic terms: results are mixed and depend heavily on the model and target language.

The **Source+Target Lexicon** condition was found to be the least stable among the models. Probably this happens because providing the full dictionaries for both languages introduces context-window overload, forces the model to identify the relevant entry among a large number of irrelevant ones, and can ultimately be overridden by the model's own pretraining priors.

Finally, in the worsened group ($28-33\%$ across categories), **Sense-Informed** loses reached on average -0.031 to -0.054 COMET points and in **Source+Target Lexicon** -0.025 to -0.030 COMET points. These cases typically arise when the model has used the correct offensive term and the extra information pushes it toward a less natural or register-mismatched alternative, or when strong pretraining priors override the prompt. For example, the Greek τσιγκούνης *tsigounis* "stingy" into Italian the **Baseline** was correctly translated *tirchio* "stingy", whereas **Sense-Informed** shifted translation to *povero* "poor", changing the meaning entirely (COMET 0.91 vs. 0.67).

6 Conclusions

This work tested whether HurtNet — the first multilingual dictionary of hurtful words — helps large language models preserve the offense load when translating between Arabic, Bulgarian, Greek, French, and Italian. We compared translations produced by three models (GPT-5.4-mini, Gemma 3 12B, MiLMMT-46-12B) under four prompt strategies: **Baseline**, **Offense-Aware**, **Sense-Informed**, and **Source+Target Lexicon**, evaluated through BLEU, chrF++, mBERT F1, and COMET.

The results confirm HurtNet's effectiveness as a resource for offensive language machine translation. Larger models benefit more from its sense-based representations, and prompts enriched with lexical knowledge consistently outperform those that merely signal offensiveness — confirming that dictionary knowledge, rather than offensive language awareness per se, drives quality improvements. The most substantial gains — for COMET often between $+0.17$ and $+0.29$ — occur where cross-lingual sense alignments are most valuable: entries lacking straightforward literal translations, such as false friends (e.g., FR *baiser* to fuck" vs IT *baciare* to kiss"), metaphorical uses (e.g., EL μπουρδέλο *bourdelo* brothel" used as chaos", σκουπίδι *skoupidi* "trash" as a human-directed insult), and culturally specific sexual expressions.

Machine translation is, however, only one of the many possible testbeds for HurtNet. Its real strength lies in its uniqueness in the field, representing the first multilingual dictionary specifically dedicated to offensive language, providing a unified lexical basis for cross-linguistic research. Its relevance extends beyond computational applications, offering valuable insights for linguistic, social, educational, and cultural studies of offensive discourse. The resource can serve as a basis for future expansion, both by incorporating additional languages in a unified structure and by enriching existing ones with dialectal, regional, or stylistic variants (e.g., Arabic dialects or Italian regionalisms).

Limitations

The number of HurtNet entries differed substantially among the five languages, resulting in a relatively small number of shared semantic categories, namely 31, and a correspondingly modest number of sense-related offensive terms. Even in this situation, the gold set could be more representative of the five languages by providing translations containing synonyms of offensive entries. For example, EL πηδάω *pidao* "to jump" is frequently used with the meaning "to fuck" in colloquial Greek. One of the translations produced by the models featured πηδάω *pidao* with this offensive meaning, which was fine, but the metrics penalised it because it was not in the gold set. Such a set up would probably have returned better metrics results regarding translation quality.

Ethics Statement

This work involves collecting, annotating, and processing offensive and potentially harmful language across different languages and cultural contexts.

The resource presented in this article, HurtNet, is intended solely for research purposes such as evaluating machine translation, lexicography, and content moderation.

As offensive language is highly context-dependent and varies across cultures, the inclusion of a term in HurtNet should not be interpreted as an absolute or universal definition of what is offensive.

The resource reflects specific language assessments made during the annotation and validation processes, and may not capture all sociolinguistic nuances, dialectal variations, or evolving uses of offensive language.

In particular, some terms may be perceived differently depending on the region, the speaker's identity, community norms, or the pragmatic context. Some definitions and examples were generated or assisted by large language models due to the limited availability of lexical resources for specific languages and meanings. All automatically generated content was manually reviewed and corrected by native speakers or domain experts before being included in the final resource.

The release of multilingual offensive lexicons may create risks of misuse, including the generation or amplification of harmful content. To mitigate these risks, HurtNet is distributed primarily as a linguistic and lexicographic resource, and the paper does not provide systems optimized for generating offensive language. Finally, we acknowledge that automatic evaluation metrics such as BLEU, chrF++, COMET, and BERTScore cannot fully capture the pragmatic and sociocultural dimensions of how offensiveness is preserved across languages. Human interpretation remains essential when assessing offensive meaning and its translation across linguistic communities.

References

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2018. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 69–76. IEEE.

Zhivko Angelov, Kiril Simov, Petya Osenova, and Zara Kancheva. 2022. The CLaDA-BG Dictionary Creation System: Specifics and Perspectives. In *CLARIN Annual Conference Proceedings*, pages 24–28, Prague, Czechia.

Iván Árcos and Jaime Pérez. 2023. Detecting hurtful humour on twitter using fine-tuned transformers and 1d convolutional neural networks. In *IberLEF@SEPLN*.

B. T. S. Atkins and M. Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford University Press, Oxford.

José Javier Ávila-Cabrera. 2023. Once upon a time in hollywood: Analysis of dubbed and subtitled insults into european spanish. *Journal of Pragmatics*.

Valerio Basile and Christian Cagnazzo. 2021. Litescale: A lightweight tool for best-worst scaling annotation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 121–127, Held Online. INCOMA Ltd.

Valerio Basile, Stella Markantonatou, Vivian Stamou, Iakovi Alexiou, Georgia Apostolopoulou, Vasiliki Archonti, Antonis Balas, Eleni Koutli, and Maria Panagiotopoulou. 2023. Using hurtlex and best–worst scaling to develop eris: A lexicon for offensive language detection. In Elías Said-Hung and Julio Montero-Díaz, editors, *News Media and Hate Speech Promotion in Mediterranean Countries*, pages 32–49. IGI Global.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018*, page 51–56. Accademia University Press.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

CIDOC CRM Special Interest Group. 2021. *CIDOC Conceptual Reference Model (CRM), Version 7.2*. ISO 21127:2023 correspondence document.

Marta R. Costa-jussà, Eric Smith, Christophe Ropers, Daniel Licht, Jean Maillard, Javier Ferrando, and Carlos Escolano. 2023. Toxicity in multilingual machine translation at scale. *Preprint*, arXiv:2210.03070.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Tullio De Mauro. 2016a. Il nuovo de mauro.

Tullio De Mauro. 2016b. Le parole per ferire. *Internazionale*. Compiled for the "Joe Cox" Committee on intolerance, xenophobia, racism and hate phenomena of the Italian Chamber of Deputies, which issued a Final Report in 2017.

Alphaeus Dmonte, Shrey Satapara, Rehab Alsudais, Tharindu Ranasinghe, and Marcos Zampieri. 2025. On the effects of machine translation on offensive language detection. *Social Network Analysis and Mining*, 14(1).

Bradley Efron. 1987. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185.

W Ferroggiaro. 2018. Social media and conflict in nigeria: A lexicon of hate speech terms.

Luca Giordano and Maria Pia Di Buono. 2023. Assessing Italian news reliability in the health domain through text analysis of headlines. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 538–548, Vienna, Austria. NOVA CLUNL, Portugal.

Levon Haroutunian. 2022. Ethical considerations for low-resourced machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 44–54, Dublin, Ireland. Association for Computational Linguistics.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

Adam Kilgarriff, Miloš Husák, Katy McAdam, Michael Rundell, and Pavel Rychlý. 2008. GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the 13th EURALEX International Congress*, pages 425–432, Barcelona, Spain.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43, Online. Association for Computational Linguistics.

Stella Markantonatou, Vivian Stamou, Christina Christodoulou, Georgia Apostolopoulou, Antonis Balas, and George Ioannakis. 2024. The corpus AIKIA: Using ranking annotation for offensive language detection in Modern Greek. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15861–15871, Torino, Italia. ELRA and ICCL.

Ayme Arango Monnar, Jorge Perez Rojas, and Barbara Poblete Labra. 2024. Cross-lingual hate speech detection using domain-specific word embeddings. *PLOS ONE*, 19(7):e0306521.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on arabic social media. In *First Workshop on Abusive Language Online 2017*, pages 52–56. Association for Computational Linguistics (ACL).

Arianna Muti, Federico Ruggeri, Cagri Toraman, Lorenzo Musetti, Samuel Algherini, Silvia Ronchi, Gianmarco Saretto, Caterina Zapparoli, and Alberto Barrón-Cedeño. 2024. Pejorativity: Disambiguating pejorative epithets to improve misogyny detection in italian tweets. *arXiv preprint*.

Ari Muzakir, Kusworo Adi, and Retno Kusumaningrum. 2023. Advancements in semantic expansion techniques for short text classification and hate speech detection. *Ingénierie des systèmes d information*, 28(3):545–556.

Petya Osenova. 2024. On a hurtlex resource for Bulgarian. In *Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria (CLIB 2024)*, pages 214–219, Sofia, Bulgaria. Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences.

Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370, Florence, Italy. Association for Computational Linguistics.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477 – 523.

Francesco Sabatini and Vittorio Coletti. 2018. Il sabatini-coletti - dizionario della lingua italiana.

Yuzhe Shang, Pengzhi Gao, Wei Liu, Jian Luan, and Jinsong Su. 2026. Scaling model and data for multilingual machine translation with open large language models. *Preprint*, arXiv:2602.11961.

Kiril Simov and Petya Osenova. 2023. Recent developments in BTB-WordNet. In *Proceedings of the 12th Global Wordnet Conference*, pages 220–227, University of the Basque Country, Donostia - San Sebastian, Basque Country. Global Wordnet Association.

Vivian Stamou, Iakovi Alexiou, Antigone Klimi, Eleftheria Molou, Alexandra Saivanidou, and Stella Markantonatou. 2022. Cleansing & expanding the HURTLEX(el) with a multidimensional categorization of offensive words. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 102–108, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Ranka Stanković, Jelena Mitrović, Danka Jokić, and Cvetana Krstev. 2020. Multi-word expressions for abusive speech detection in Serbian. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 74–84, online. Association for Computational Linguistics.

Jherez Taylor, Melvyn Peignon, and Yi-Shin Chen. 2017. Surfacing contextual hate speech words within social media. *Preprint*, arXiv:1711.10093.

Gemma Team. 2025. Gemma 3.

Alice Tontodimamma, Lara Fontanella, Stefano Anzani, and Valerio Basile. 2023. An italian lexical resource for incivility detection in online discourses. *Quality & Quantity: International Journal of Methodology*, 57(4):3019–3037.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language.

Jinfang Yao, Shaidatul Akma Adi Kasuma, and Hisham Noori Hussain Al-Hashimy. 2025. Exploring ethical considerations in machine translation: Addressing bias and cultural sensitivity. *Journal of Interdisciplinary Studies in Education*, 14(4):99–116.

Xiaojing Zhao, Emmanuele Chersoni, Chu-Ren Huang, and Han Xu. 2026. How do language models handle emotional content in video game localization? a computational linguistics approach. *Research Methods in Applied Linguistics*, 5(1):100294.

A Prompt Conditions

Baseline Condition Prompt

```
Translate the source sentence into the target language.

Return ONLY valid JSON with exactly these two fields:

{
"Sentence_Target": "...",
"Reason": "..."
}

Rules:
- "Sentence_Target" must contain only the translated sentence in the target language.
- "Reason" must be written in English.
- "Reason" must briefly explain the translation choice in 1-3 sentences.
- Do not add notes outside the JSON.
- Do not add markdown.
- Do not write anything before or after the JSON.

Source language: {source_lang}
Target language: {target_lang}
Source sentence: {source_sentence}
```

Offensive-Aware Condition Prompt

```
You are a professional machine translation system. Translate the source sentence into the target language.
The source sentence contains an offensive expression.
Translate the sentence naturally into the target language without losing the offensive meaning.
Preserve the offensive force and function as much as possible, but make the translation natural in the target language.

Return ONLY valid JSON with exactly these two fields:

{
"Sentence_Target": "...",
"Reason": "..."
}

Rules:
- "Sentence_Target" must contain only the translated sentence in the target language.
- "Reason" must be written in English.
- "Reason" must briefly explain the translation choice in 1-3 sentences.
- Do not add notes outside the JSON.
- Do not add markdown.
- Do not write anything before or after the JSON.

Source language: {source_lang}
Target language: {target_lang}
Source sentence: {source_sentence}
```

Sense-Informed Condition Prompt

```
You are a professional machine translation system. Translate the source sentence into the target language. The source sentence contains an offensive expression. Use the sense-level alignment information below to preserve the offensive meaning. Choose the aligned target-language expression that best matches the meaning, pragmatic function, target, and offensive force of the source expression in this sentence. Translate the whole sentence naturally into the target language.

Return ONLY valid JSON with exactly these two fields:

{
"Sentence_Target": "...",
"Reason": "..."
}

Rules:
- "Sentence_Target" must contain only the translated sentence in the target language.
- "Reason" must be written in English.
- "Reason" must briefly explain the translation choice in 1-3 sentences.
- Do not add notes outside the JSON.
- Do not add markdown.
- Do not write anything before or after the JSON.

Source language: {source_lang}
Target language: {target_lang}
Source offensive expression: {source_lemma}
Source definition: {source_definition}
Possible aligned {target_lang} offensive expressions with original-language definitions: {candidate_block}
Source sentence: {source_sentence}
```

Source+Target Lexicon Condition Prompt

```
You are a professional machine translation system. Translate the source sentence into the target language. The source sentence contains an offensive expression. You are given two offensive-language lexicons:

1. the full source-language offensive lexicon,
2. the full target-language offensive lexicon.
First, infer from the source sentence which source-language lexicon entry best matches the offensive expression and its sense. Then, choose from the target-language lexicon the expression that best preserves the same meaning, pragmatic function, target, and offensive force. Translate the whole sentence naturally into the target language without losing the offensive meaning.

Return ONLY valid JSON with exactly these two fields:

{
"Sentence_Target": "...",
"Reason": "..."
}

Rules:
- "Sentence_Target" must contain only the translated sentence in the target language.
- "Reason" must be written in English.
- "Reason" must briefly explain the translation choice in 1-3 sentences.
- Do not add notes outside the JSON.
- Do not add markdown.
- Do not write anything before or after the JSON.

Source language: {source_lang}
Target language: {target_lang}
Full {source_lang} offensive-language lexicon: {source_lexicon_block}
Full {target_lang} offensive-language lexicon: {target_lexicon_block}
Source sentence: {source_sentence}
```

B Analysis of Per-Language-Pair

The plots report the delta scores — that is, the difference in metric value relative to the **Baseline** condition — achieved by the other three prompting strategies (**Source+Target Lexicon**, **Offense-Aware**, and **Sense-Informed**) for GPT-5.4-mini, the best-performing model overall. Results are shown across 20 translation directions involving five languages: Arabic, Bulgarian, French, Greek, and Italian. Each direction is evaluated under complementary metrics: BERTScore (a contextual embedding-based similarity measure), chrF++ (character-level overlap metric accounting for word order), and COMET (a neural reference-based met-

ric trained on human judgments). A green bar indicates that the strategy improves over the **Baseline** for that direction and metric, while a red bar indicates degradation.

Figure 2: Per-language-pair delta scores relative to the **Baseline** for GPT-5.4-mini

Figure 2 shows the BertScore results: the **Sense-Informed** condition tends to yield the largest positive gains, particularly for directions involving Greek or Italian as the target language (Italian→Greek reaches +0.020, the highest value in the graph) and Arabic→French. The **Offense-Aware** strategy is the most inconsistent: it often produces negative deltas, especially when Arabic is the target language or Greek is the source language (Bulgarian→Arabic; Greek→Arabic/Bulgarian). The **Source+Target Lexicon** strategy is conservative—modest or near-zero positive deltas in most cases; the worst result is recorded with Bulgarian→Arabic, the best with Arabic→French. Translations into Arabic and from Greek are overall the most problematic in terms of the semantic similarity of the translations.

As for the chrF++ metric, the results in Figure 3 suggest that **Sense-Informed** tends to perform relatively well, with notable positive deltas observed for Arabic→Italian, Arabic→French, and Italian→Greek. On the other hand, Bulgarian→Arabic and French→Arabic display negative deltas for both **Source+Target Lexicon** and **Sense-Informed**, which confirm the complexity of Arabic as a target language. **Offense-Aware**

Figure 3: Per-language-pair delta scores relative to the **Baseline** for GPT-5.4-mini

generally underperforms in this metric as well, with negative deltas appearing across several directions including Greek→Arabic, Greek→Bulgarian, Greek→French, and French→Arabic further highlighting the need for additional information in addition to the flag indicating the presence of offensive content.

Figure 4: Per-language-pair delta scores relative to the **Baseline** for GPT-5.4-mini

As for COMET in Figure 4, negative deltas

are far more prevalent across all strategies and language pairs than in any other metric. Greek→Arabic is consistently the worst direction (all three bars deeply negative). French→Arabic **Offense-Aware** drops to nearly -0.020, the largest single negative in the whole set. Arabic→Italian is an exception, showing positive deltas for **Source+Target Lexicon** and **Offense-Aware**. **Sense-Informed** does not dominate here as clearly as in the previous metrics; in fact, it sometimes produces the worst COMET score (Bulgarian→Italian; Greek→Italian). This divergence between COMET and the surface metrics (BLEU/chrF++) is meaningful: strategies that improve lexical overlap may still degrade fluency and adequacy as judged by a neural metric.

C Scores by Source and Target Language

The following two tables summarise the full automatic evaluation results. The first table groups the results by source language; the second groups them by target language. BLEU and chrF++ are on a 0–100 scale, while mBERT F1 and COMET are on a 0–1 scale.

Table 5: Overall automatic scores by Source Language.

Source Language	n	Mean BLEU	Mean chrF++	Mean mBERT F1	Mean COMET
Arabic	1920	27.59	55.03	0.8354	0.8645
Bulgarian	1920	27.65	54.93	0.8452	0.8617
Greek	1920	27.83	53.30	0.8419	0.8563
French	1920	28.91	**55.34**	0.8472	**0.8806**
Italian	1920	**29.43**	55.32	**0.8500**	0.8749

Table 6: Overall automatic scores by Target Language.

Target Language	n	Mean BLEU	Mean chrF++	Mean mBERT F1	Mean COMET
Arabic	1920	18.93	45.23	0.8430	0.8368
Bulgarian	1920	33.18	58.42	**0.8477**	**0.8994**
Greek	1920	25.45	53.29	0.8375	0.8928
French	1920	30.16	57.66	0.8473	0.8485
Italian	1920	**33.70**	**59.31**	0.8443	0.8605

D Further Statistical Validation of Prompting Condition Effects

To assess whether the differences in translation quality across prompting conditions are statistically reliable, we conducted a further analysis on the per-direction scores (20 language pairs per condition per model) for each of the four evaluation metrics (BLEU, chrF++, mBERT F1, COMET).

Friedman test We applied a non-parametric Friedman test, treating language pair as the blocking factor and prompting condition as the treatment, separately for each model and metric (12 tests in total). Results show that condition effects are significant for Gemma-3 12B across all four metrics ($\chi^2 > 25$, all $p < .0001$). For MiLMMT-46 12B and GPT-5.4, significance is reached only on mBERT F1 ($\chi^2 = 10.27$, $p = .016$ and $\chi^2 = 18.78$, $p = .0003$, respectively). BLEU, chrF++, and COMET show no significant variation across conditions for the two larger models, indicating that their overall translation quality remains stable while lexical choices vary.

Pairwise Wilcoxon signed-rank tests We compared each non-baseline condition against the **Baseline** using paired Wilcoxon signed-rank tests (pairing on language pair), with Holm correction applied across all 36 comparisons (3 conditions $\times$ 4 metrics $\times$ 3 models). After correction, five contrasts are significant ($p < .05$): the **Offense-Aware** condition significantly degrades Gemma-3 12B on all four metrics ($p = .003$); the **Source+Target Lexicon** condition degrades Gemma-3 12B on BLEU ($p = .032$); and **Sense-Informed** prompting yields a significant positive effect for GPT-5.4-mini on mBERT F1 ($p = .031$).

Bootstrap confidence intervals To quantify uncertainty around the mean delta scores and to detect consistent effects that the Holm-corrected Wilcoxon may miss due to limited statistical power ($n = 20$ blocks), we computed bias-corrected and accelerated (BCa) bootstrap confidence intervals (Efron, 1987) at the 95% level ($R = 5{,}000$ resamples) on per-direction deltas relative to the **Baseline**. An effect is flagged as bootstrap-significant (†) when the CI excludes zero but the Wilcoxon test does not reach the corrected threshold — indicating a consistent directional effect whose formal detection is limited by sample size and multiple-comparison adjustment.

Linear mixed models As a complementary analysis that accounts for the full covariance structure, we fitted one linear mixed model (LMM) per metric, with prompting condition and model as fixed effects and language pair as a random intercept (fitted by maximum likelihood). Random-effect variance substantially exceeded residual variance across all metrics (ratio $\approx$ 8:1 for BLEU and chrF++), confirming that translation direction is the dominant source of variability. Among fixed effects, **Offense-Aware** was the only condition with a significant negative effect across all four metrics (BLEU: $\beta = -1.46$, $t = -3.74$; chrF++:

Table 7: Mean delta scores relative to **Baseline**, 95% BCa bootstrap confidence intervals, and Holm-corrected Wilcoxon p-values for all prompting conditions, models, and metrics. Significance codes: $^{**}p < .01$, $^{*}p < .05$ (Wilcoxon, Holm-corrected); † 95% BCa CI excludes zero (Wilcoxon $p \geq .05$ after correction). Positive deltas indicate improvement over **Baseline**.

		BLEU			chrF++			mBERT F1			COMET		
Model	**Condition**	Δ	95% CI	p	Δ	95% CI	p	Δ	95% CI	p	Δ	95% CI	p
Gemma-3 12B	Offense-Aware	-4.24^{**}	$[-5.48, -3.14]$	.004	-2.65^{**}	$[-3.60, -1.85]$	.003	-0.015^{**}	$[-0.018, -0.012]$	.003	-0.037^{**}	$[-0.043, -0.031]$	.003
	Sense-Informed	-0.49	$[-1.40, +0.53]$	1.00	-0.01	$[-0.93, +0.85]$	1.00	$+0.002$	$[-0.002, +0.005]$	1.00	$-0.006^{\dagger}$	$[-0.011, -0.001]$	1.00
	Source+Target Lexicon	-2.03^{*}	$[-2.89, -1.17]$	.032	$-1.65^{\dagger}$	$[-2.37, -0.88]$	.062	$-0.005^{\dagger}$	$[-0.008, -0.003]$	.062	-0.003	$[-0.007, +0.001]$	1.00
MiLMMT-46 12B	Offense-Aware	-0.33	$[-0.66, +0.20]$	.908	-0.29	$[-0.68, +0.03]$	1.00	-0.001	$[-0.002, +0.001]$	1.00	-0.002	$[-0.003, +0.000]$	1.00
	Sense-Informed	$+0.34$	$[-0.19, +0.83]$	1.00	-0.11	$[-0.59, +0.42]$	1.00	$+0.001$	$[-0.001, +0.002]$	1.00	$+0.001$	$[-0.001, +0.003]$	1.00
	Source+Target Lexicon	$+0.47$	$[-0.23, +1.10]$	1.00	-0.18	$[-0.85, +0.35]$	1.00	$+0.003^{\dagger}$	$[+0.001, +0.004]$	.143	$+0.001$	$[-0.001, +0.004]$	1.00
GPT-5.4-mini	Offense-Aware	$+0.19$	$[-0.55, +1.03]$	1.00	$+0.14$	$[-0.63, +1.05]$	1.00	-0.001	$[-0.004, +0.002]$	1.00	-0.003	$[-0.007, +0.001]$	1.00
	Sense-Informed	$+1.58^{\dagger}$	$[+0.55, +2.72]$	.585	$+0.37$	$[-0.48, +1.25]$	1.00	$+0.007^{*}$	$[+0.004, +0.010]$	.031	-0.001	$[-0.005, +0.002]$	1.00
	Source+Target Lexicon	$+0.55$	$[-0.15, +1.27]$	1.00	-0.19	$[-0.96, +0.49]$	1.00	$+0.001$	$[-0.001, +0.004]$	1.00	$+0.000$	$[-0.003, +0.004]$	1.00

$\beta = -0.93$, $t = -2.93$; mBERT F1: $\beta = -0.006$, $t = -4.70$; COMET: $\beta = -0.014$, $t = -7.36$; all $p < .05$, Holm-corrected Kenward–Roger contrasts). **Sense-Informed** showed a significant positive effect on mBERT F1 only ($\beta = +0.003$, $t = 2.59$, $p = .021$), while **Source+Target Lexicon** did not reach significance on any metric.

Summary of results Table 7 reports mean delta scores, 95% BCa bootstrap confidence intervals, and Holm-corrected Wilcoxon p-values for all 36 condition–model–metric combinations. Three findings emerge consistently across all four analytical stages.

First, **Offense-Aware** prompting reliably degrades Gemma-3 12B across all metrics, with large negative deltas (ΔBLEU $= -4.24$, 95% CI $[-5.48, -3.14]$; ΔCOMET $= -0.037$, 95% CI $[-0.043, -0.031]$; $p = .003$). This confirms that flagging offensive content without providing structured lexical guidance destabilises smaller models, consistently with the qualitative observations in Section 5.

Second, **Sense-Informed** prompting is the only condition to produce significant positive effects, but exclusively for GPT-5.4-mini on mBERT F1 ($\Delta = +0.007$, 95% CI $[+0.004, +0.010]$, $p = .031$) and BLEU ($\Delta = +1.58$, 95% CI $[+0.55, +2.72]$, $\dagger$). This pattern indicates that dictionary knowledge benefits larger general-purpose models by improving offensive-word selection, without destabilising fluency or adequacy as measured by COMET.

Third, **Source+Target Lexicon** produces no significant positive effects for any model after correction. For Gemma-3 12B it yields significant degradation on BLEU ($p = .032$) and bootstrap-significant degradation on chrF++ and mBERT F1, corroborating the hypothesis that providing full dictionaries for both languages introduces context overload and is overridden by the model's pretraining priors. A small but consistent mBERT F1 gain is detected for MiLMMT-46 12B by the bootstrap ($\Delta = +0.003$, 95% CI $[+0.001, +0.004]$, $\dagger$), suggesting marginal benefit from full-lexicon injection for the translation-dedicated model.