**Frontiers in Computer Science and Artificial Intelligence**
ISSN: 2978-8048
DOI: 10.32996/fcsai
Journal Homepage: www.al-kindipublisher.com/index.php/fcsai

FCSAI
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

**| RESEARCH ARTICLE**

# Decoding Digital Evasion: Recognition and Interpretation of Obfuscated Arabic Text in Images by Gemini, Copilot, Claude, and Qwen

**Reima Al-Jarf**
*Full Professor of English and Translation Studies, Riyadh, Saudi Arabia*
**Corresponding Author**: Reima Al-Jarf, **E-mail:** reima.al.jarf@gmail.com

## | ABSTRACT
This study examines how four AI models—Gemini, Copilot, Claude, and Qwen 3.7 Plus—decode and interpret obfuscated Arabic words embedded in 138 images extracted from YouTube and Instagram videos posted during the USA–Israel–Iran War (March–April 2026). The obfuscation relied primarily on inserting punctuation marks (periods, commas, dashes, underscores) to evade platform moderation. Across these images, the dataset contained 374 obfuscated Arabic words, which the models translated in context and in isolation. Analysis revealed variations in recognition and interpretation across the models. All four models correctly translated 30% of images; three models succeeded in 24%; and two models converged on 27%. Gemini produced uniquely correct translations for 16% of the images. All four models failed completely on 3% of the images, where root semantics were unrecoverable. Additionally, Gemini translated Arabic text in 94% of the images correctly, producing accurate translations for 98% of the words in context and 99% in isolation. Copilot succeeded on 61% of the images, correctly translating 72% in context and 98% in isolation. Qwen translated 54% of the images, yielding correct translations for 72% of the words in context and 98% in isolation. Claude also succeeded on 54% of the images, producing 67% correct words in context and 98% in isolation. Across all models, recognition accuracy was consistently higher on isolated words than within full images, indicating that failures were driven by visual noise, contextual interference, and safety-filter activation, rather than linguistic limitations. Additional behaviors included deleted and replaced tokens, hallucinated translations, transliteration, and partial clipping. Copilot blocked few images or produced gibberish output. Gemini sometimes mirrored obfuscation patterns in English equivalents. In isolated-word translation, the models produced identical and synonymous equivalents. Claude added feminine gender markers to some distorted words. The lexemes exhibiting the highest number of polymorphic distortion variants were Iran, Israel, war, missiles, nuclear, America, attack, Netanyahu, party/war, Gaza and Zionist. How LLMs process obfuscated Arabic text, density, complexity, and symbol typology of obfuscation, contextual vs. isolated processing, the spectrum of mimetic obfuscation, non-obfuscation hallucination, algorithmic paranoia, keyword panic, and socio-cultural misalignment are discussed in detail.

## | KEYWORDS
Linguistic encryption, obfuscated Arabic words, distorted text recognition, algospeak, algorithmic moderation, content moderation, multimodal AI models, OCR segmentation, political lexeme distortion, social media images.

## | ARTICLE INFORMATION

**ACCEPTED:** 02 June 2026 **PUBLISHED:** 11 July 2026 **DOI:** 10.32996/fcsai.2026.5.9.11

## 1. Introduction
Linguistic encryption[1] refers to the intentional use of phonetic or orthographic alterations, codes, or playful language to hide or mask meaning. Unlike technical encryption, which relies on mathematical principles and cryptographic ciphers, linguistic encryption relies on cultural knowledge, wordplay, and context. Linguistic encryption is also called Algospeak[2] (algorithm + speak) which refers to coded language, like new words, misspellings, or emojis, used on social media to bypass automated content moderation systems that flag and remove sensitive topics (politics, sex, self-harm). Algospeak serves as a strategic tool for users to discuss taboo or restricted subjects without risking "shadow-banning" or de-platforming. Consequently, it has evolved into a highly dynamic digital dialect where terms mutate rapidly to outpace algorithmic updates.

[1] https://www.naukri.com/blog/what-is-content-moderation/
[2] https://englishlanguagestudies.com/algospeak-social-media-driven-language/

Linguistic encryption and Algospeak use several strategies/techniques[3], including: (i) phonetic distortion and masking, where sounds are slightly altered to mock or disguise serious names; (ii) orthographic manipulation, utilizing letter substitution, deletion, mixed scripts, or transliteration that obscures meaning; (iii) semantic downgrading, trivializing powerful institutions or military systems by replacing them with fragile or silly imagery; (iv) animal metaphors, employing dehumanizing comparisons to ridicule public figures; (v) food-based encryption, mocking military units or political factions through culinary references; (vi) institutional satire, renaming bureaucracies or organizations with absurd labels; (vii) encrypted insults and slurs, which disguise derogatory references under coded language; (viii) invented distortions and neologisms, coining words or spellings that sound nonsensical but carry insider meaning; (ix) satirical circumlocutions and figurative metaphors, cloaking direct references in creative imagery; (x) political and geographic distortions, deliberately twisting place names, leaders, or geopolitical groups; (xi) Lexical scattering, i.e., inserting coded words into otherwise normal sentences so only insiders catch the reference; (xii) cultural anchoring, relying on insider cultural references; and (xiii) religious and cultural invocations, mixing sacred or cultural language with satire (Al-Jarf, 2025e; Al-Jarf, 2025k; Al-Jarf, 2023a; Al-Jarf, 2023b; Al-Jarf, 2022a; Al-Jarf, 2022b; Al-Jarf, 2021c; Al-Jarf, 2011; Al-Jarf, 2010; Al-Jarf, 1998; Al-Jarf, 1995). Another strategy for distorting written words in Arabic involves inserting symbols and punctuation marks between the letters of a word or replacing Arabic characters with numbers, symbols or English letters.

While humans easily decode these encrypted meanings by relying on shared culture, context, and tone, social media algorithms often struggle due to their reliance on surface-level keyword detection. Because scattered slang embedded within natural speech patterns does not match static lists of banned terms, users successfully construct a "masked lexicon." This lexicon communicates freely with the intended audience while effectively bypassing algorithmic moderation systems.

Given the widespread adoption of social media communication and increasingly strict algorithmic moderation, a review of the literature has revealed extensive research that examined algorithmic moderation, platform governance, and legal frameworks on social media platforms. Prior studies can be organized into three distinct groups. The first group of studies investigated algorithmic moderation, platform governance, and legal frameworks. These studies include how pro-Palestine activists navigate algorithmic visibility (Hess & Häusermann, 2026); the metamorphosis of censorship in the digital age (Díaz Cuesta & Barros Garbín, 2026); the meaning, process, and role of moderators (Naukri Content Team, 2025), how Chinese social media platforms moderate content (Li & Zhou, 2024; ); critique of algorithmic arbitrariness in content moderation (Gomez, 2024); algorithmic enforcement (Prem & Krenn, 2024), TikTok and Palestinian narratives as an example of algorithmic imperialism and digital resistance (Abbassa, 2023); content moderation in social media: from guidelines to enforcement (Singhal, 2023); using algospeak to contest algorithmic moderation on TikTok (Steen, 2023); and technological dimensions of moderation (Liu, Yildirim & Zhang, 2021). Other studies contextualized moderation within broader social and informational frameworks (Morrow et al., 2022),and and assessed current challenges and future directions (Gongane, Munot & Anuse, 2022). Scholars have also set enforcement guidelines and practices (Singhal et al., 2023); identified detection and moderation of detrimental content on social media platforms (Gongane, 2022); contextualized social media moderation (Morrow, 2022); the difficulties of social media content moderation (Young, 2022); algorithmic content moderation on social media in EU law (Castets-Renard, 2020), user experiences contesting moderation on Facebook (Vaccaro, et al., 2020) and moderation techniques for social media (Veglis, 2014). Collectively, these studies demonstrate the complexity of the current moderation systems and their limitations in detecting coded or context-dependent language.

The second group of studies explored data encryption, security, and detection of obscured data. These studies include generative AI and difficulties of conventional encryption (Nandhini, 2026); AI-enhanced AES encryption for Kurdish unicode texts (Abduljabbar, 2026); word-based encryption algorithms using dictionary indexing with variable key length (Al-Jarrah et al., 2022); and reliable detection of compressed and encrypted data (De Gaspari et al., 2022).

A third group of studies explored user evasion strategies, algospeak, and large language model (LLM) decoding such as multimodal strategies and platform governance (Elyamany, 2026); the trade-off between legible meaning of algospeak and detection avoidance (Fillies, 2026); how Jihadist supporters exploit algorithmic recommendations and evade content moderation on Tiktok (Karo, 2026); censorship of TikTok algospeak on social media (Moskal & Supernak, 2026); Multimodal communication manipulation in response to algorithmic content moderation on TikTok (Young, 2026); decoding algospeak in Chinese social media captions (Zhao, 2026); content creators' understandings of TikTok's algorithms (Florini et al., 2026); algospeak and sensitive language on TikTok (Blažević, Žuvela & Blažević, 2025); how algorithm awareness impacts algospeak use on TikTok (Felaco & Pelliccia, 2024; Klug, et al., 2023); how algospeak is transforming the future of language on social media (Aleksic, 2025); how community expertise moderates toxic algospeak (Agarwal & Falade, 2026); Algospeak on social media platforms with Orange-based text-mining (Pocock, 2025); algorithmic hate speech detection in online communication between adolescents (Fillies, 2025); algospeak and communication styles on social media in the era of AI (Isam, 2025); algospeak and digital culture as social media challenges (Isam, 2024); how algospeak is changing our language in real time (Lorenz, 2022); cryptic language through ChatGPT (Youvan, 2023); can AI decode and interpret encrypted Arabic on Facebook and YouTube to evade algorithmic moderation (Al-Jarf, 2025e); and others.

[3] https://www.naukri.com/blog/what-is-content-moderation/

Although prior studies have extensively examined algorithmic moderation, platform governance, and English-based algospeak, focus on TikTok or Facebook, analyzes English or Chinese coded language, and address technical ciphers rather than culturally grounded linguistic distortion, there is a dearth of studies on algospeak, encrypted words and algorithmic moderation in Arabic. Only one study by Al-Jarf 2025) examined the ability of three AI models (DeepSeek, Copilot, and Google Translate) to decode and interpret spoken encrypted Arabic political expressions from YouTube videos during the 2024–2025 Gaza–Israel war, as well as written encrypted Arabic words and expressions related to COVID-19 from Facebook posts during the 2021–2023 pandemic. The encrypted expressions in that study revealed phonetic distortion, orthographic manipulation (letter substitution, deletion, mixed scripts, transliteration), semantic downgrading, animal metaphors, food-based encryption, institutional satire, encrypted insults, invented distortions, neologisms, satirical circumlocutions, figurative metaphors, political and geographic distortions, cultural anchoring, and religious or cultural invocations. However, Al-Jarf (2025) did not examine obfuscated Arabic words—visually distorted forms embedded inside images, nor did it address OCR limitations or safety-layer interventions. To date, no comparative study has evaluated modern generative AI models' ability to detect, decode, or interpret obfuscated Arabic words on YouTube or Instagram, and no research has analyzed how OCR errors and safety-layer constraints shape model performance. This gap is especially salient in political contexts where Arab users intentionally distort language to evade algorithmic moderation.

Therefore, the present study seeks to fill this gap by investigating whether AI models can decode and interpret obfuscated Arabic words—both in context and in isolation—embedded in images taken from YouTube and Instagram videos created by Arab digital users during the USA–Israel–Iran War in March and April 2026 to avoid detection by platform moderation systems. The study compares the performance of four generative AI (GenAI) models: Gemini, Copilot, Claude and Qwen3.7-Plus in decoding and interpreting obfuscated Arabic words in context as they appear in the sample images and in isolation. This study aims to answer the following questions: (i) To what extent can generative AI models (Gemini, Copilot, Claude, and Qwen) detect Arabic obfuscated words embedded in YouTube and Instagram images? (ii) How accurately can these AI models decode and interpret the meanings of obfuscated Arabic words, both in isolation and in contextualized political content? (iii) How do OCR limitations and safety-layer interventions influence each model's ability to recognize and interpret obfuscated Arabic text? (iv) What differences exist among the four AI models in their detection, decoding, and interpretation performance across repeated trials?

This study is significant for several reasons. First, it addresses a major gap in the literature by examining Arabic obfuscated words, a form of linguistic distortion that has not been systematically analyzed in prior research on algospeak or content moderation. While existing studies focus primarily on English or Chinese coded language, the present study extends the discussion to Arabic, a language whose orthographic complexity, connected script, and shape-shifting letters pose unique challenges for AI detection and OCR systems.

Second, the study contributes to the emerging field of multimodal moderation research by analyzing obfuscated Arabic words embedded inside social-media images rather than plain text. This is a critical advancement because most moderation systems—and most academic studies—evaluate text typed directly into platforms, overlooking the growing use of images as a means of coded political communication. By focusing on images taken from YouTube and Instagram videos, the study reflects real user practices and real moderation vulnerabilities.

Third, the study provides the first comparative evaluation of four generative AI models—Google Gemini, Microsoft Copilot, Claude, and Qwen3.7-Plus—in detecting, decoding, and interpreting obfuscated Arabic words. This comparative dimension is essential for understanding how different AI architectures, OCR pipelines, and safety-layer policies influence model performance. The findings offer practical insights into the strengths and limitations of current AI systems and highlight inconsistencies across repeated trials, which have direct implications for platform governance and automated moderation.

Fourth, the study is significant because it examines obfuscated Arabic used in political content during the USA–Iran war (March–April 2026). This situates the research within a real geopolitical context where Arab digital creators intentionally distort words to evade algorithmic moderation. Understanding how AI models respond to politically sensitive obfuscation is crucial for improving moderation accuracy while safeguarding freedom of expression.

Fifth, the study contributes to both computational linguistics and digital-media governance by demonstrating that current AI systems still rely heavily on surface-level keyword detection and struggle with culturally grounded, visually distorted Arabic forms. The results underscore the need for more robust multimodal moderation tools capable of deeper semantic understanding, especially in languages with complex orthography.

Finally, this study is part of a series of studies by the author on the use of AI in Arabic language contexts such as a comparative study of 3 AI models in decoding and Interpreting encrypted Arabic on Facebook and YouTube to evade algorithmic moderation (Al-Jarf, 2025e); human vs AI translation of common names of chemical compounds: A comparative study (Al-Jarf, 2025k); AI and students' translation of Arabic expressions of impossibility (Al-Jarf, 2025q). It also adds to a series of studies on AI as the translation of medical terms (Al-Jarf, 2024b; Al-Jarf, 2024c); Arabic folk medical terms with *om* and *abu* (Al-Jarf, 2025q), denotative and metonymic *abu* and *umm* animal and plant folk names (Al-Jarf, 2025f); technical terms (Al-Jarf, 2021a; Al-Jarf, 2016); Arabic *abu*

brand names with different prompts (Al-Jarf, 2025d); sleep terms and formulaic expressions (Al-Jarf, 2025s); zero expressions (Al-Jarf, 2025t); Arabic grammatical terms used metaphorically (Al-Jarf, 2025i); Gaza–Israel war terminology (Al-Jarf, 2025b); full-text Arabic research articles with educational polysemes (Al-Jarf, 2025a); editors and publishers' views on the publication of AI-generated research articles in scholarly journals (Al-Jarf, 2025p; Al-Jarf, 2026); Arab instructors' views on students' assignments and research papers generated by AI (Al-Jarf, 2024a); pronunciation errors in Arabic YouTube videos narrated by AI (Al-Jarf, 2025g; Al-Jarf, 2025m; Al-Jarf, 2025n); and Arabic transliteration of borrowed English nouns with /g/ by AI (Al-Jarf, 2025c); AI translation of the Gaza-Israel war terminology (Al-Jarf, 2025b); Gemini, Ernie ViLG and Google Translate's recognition of calligraphic text in Arabic, Japanese and Chinese (Al-Jarf, 2025d); can Artificial Intelligence (AI) translate Arabic abu-brand names with different prompts (Al-Jarf, 2025f); Google translate then and now: translations from five languages into English and Arabic (2012–2025) (Al-Jarf, 2025j); Copilot vs DeepSeek's translation of denotative and metonymic abu- and umm- animal and plant folk names in Arabic (Al-Jarf, 2025h); Copilot's English Translation of Contrastive Emphatic Negation in Arabic Discourse: An Analytical Study (Al-Jarf, 2025); Specific linguistic questions that Artificial Intelligence (AI) cannot answer accurately (Al-Jarf, R, 2025o).

Additionally, the current study is part of a series of research articles that analyze Arabic political discourse including: metaphorical political slurs in Arab social media discourse describing Middle East conflicts (Al-Jarf, 2025l); political (in)correctness and the cancel-culture attitude in religious sectarian language after the Arab Spring (Al-Jarf, 2023b); emerging political expressions in Arab Spring media (Al-Jarf, 2022; Al-Jarf, 2015); sectarian language & perception of the "Other" after the Arab Spring (Al-Jarf, 2022); combating the Covid-19 hate and racism speech on social media (Al-Jarf, 2021b); sectarianism after the Arab Spring (Al-Jarf, 2019);

## 2. Methodology

### *2.1 Sample of AI Models*

This study adopts a comparative, cross-script design to examine how different AI tools decode and interpret obfuscated Arabic words in 138 images. To be included in the sample, an AI system must allow the direct upload or multiple images, must represent American and Chinese companies and technologies, and must utilize different image processing, tokenization, and visual-linguistic grounding processes. Based on these criteria, Gemini, Copilot, Claude, and Qwen3.7-Plus were selected because they represent four major AI ecosystems spanning two distinct geopolitical tech hubs and four distinct multimodal architectures —Google, Microsoft, Anthropic, and Alibaba Cloud from China. These models differ in their underlying OCR pipelines, multilingual training data, and approaches to processing Arabic script in images.

Gemini reflects a natively multimodal integration of vision and language, relying on a unified architecture where visual tokens are mapped directly into the same latent space as text tokens, preserving high spatial and geometric awareness of text layouts. Copilot incorporates Microsoft's advanced visual grounding layers alongside state-of-the-art vision-language transformer stacks, which emphasize a balanced intersection between raw pixel pixel-grid parsing and strict linguistic probability. Claude emphasizes deep contextual reasoning, interpretive coherence, and a highly literal character-by character transcription protocol, making it uniquely sensitive to structural anomalies and micro-level punctuation shifts within words. Qwen3.7-Plus represents a highly specialized, multilingual, flexible architecture developed in China, featuring custom tokenization pipelines and robust vision-transformer (ViT) encoders explicitly optimized for right-to-left (RTL) scripts and complex non-Western layout configurations. This diversity provides a broad comparative framework for examining how contemporary AI systems handle Arabic text embedded in images, including visually altered or obfuscated forms.

### *2.2 Sample of Political Images with Obfuscated Arabic Words*

A sample of 138 images of political Arabic texts with obfuscated words was collected from YouTube and Instagram videos during the US-Israel-Iran War (March–April 2026). The sample comprised 76 images from Instagram and 62 from YouTube videos. To be included in the sample, the political texts taken from the YouTube videos or Instagram looped micro-videos (Reels) had to contain at least one distorted Arabic word. Whether on YouTube or Instagram, all images in the sample were taken from publicly available videos, and no private or sensitive information was used. The images are screenshots of YouTube video titles or comments, whereas the Instagram sample consisted of screenshots of looped micro-videos (Instagram Reels), which utilize a static text overlay on a motion background format. In these specimens, the obfuscated Arabic text blocks remained stationary over a looping visual background accompanied by an audio track. This specific composition allows the text to remain visible to the user for an extended duration while exploiting platform algorithms that favor short-form video distribution. All images were cropped to conceal the background image, the name of the channel, and the name of the account owner or speaker, as well as any identifying watermarks that might compromise creator anonymity. This measure was implemented to maintain rigorous research privacy ethics and ensure absolute confidentiality.

The political images in the dataset display a wide range of font styles and visual treatments that directly influence the readability of obfuscated Arabic words. The text overlays frequently employ bold, high-contrast Arabic fonts designed to capture attention in fast-moving political content, while other images use thin or semi-thin fonts that become increasingly difficult to decipher once distortion marks are added. Many images rely on common sans-serif Arabic fonts similar to Noto Sans Arabic or Arial Arabic typefaces, which are widely used in social media graphics and news-style posts. In contrast, some images incorporate decorative or stylized fonts with shadow, glow, or outline effects that introduce artificial edges around letters. These stylistic choices

complicate segmentation and recognition, especially when combined with symbol-based distortions. In several cases, Arabic letters appear alongside Latin characters inserted within the word, creating mixed-script compositions that disrupt the natural right-to-left flow of Arabic text and further challenge multimodal AI models.

The backgrounds across the 138 political images are visually diverse and often complex. Many images feature dark or low-light backgrounds—black, navy, or charcoal—that heighten contrast with bright text overlays. Others include blurred or partially obscured backgrounds derived from video frames or cropped screenshots, producing uneven textures that interfere with text clarity. A subset of images contains patterned or ornamental backgrounds, including geometric motifs or architectural designs, which introduce visual noise and reduce the legibility of distorted words. High-saturation color fields such as red, yellow, and blue are also common, reflecting the aesthetic conventions of urgent political announcements and conflict-related media. In Instagram Reels, the background is frequently animated, with the text remaining static while the underlying video loops continuously. This composition keeps the text visible for extended periods but introduces dynamic noise that complicates Optical Character Recognition (OCR) stability and multimodal text recognition.

The interaction between font characteristics and background composition plays a decisive role in determining how readable—or intentionally unreadable—obfuscated Arabic words become. Low contrast between text and background, particularly when bright text is placed over saturated or patterned visuals, reduces the visibility of punctuation-based distortions such as dots, commas, dashes, and underscores. Motion backgrounds in Instagram Reels maintain text visibility but introduce continuous visual interference that destabilizes recognition. Decorative fonts distort letter boundaries, making it difficult for models to distinguish between genuine letterforms and distorted symbols. Shadow, glow, and outline effects create multiple artificial edges around letters, amplifying the confusion caused by symbol insertion. Mixed-script distortions, where Latin characters or numerals interrupt Arabic words, disrupt the expected orthographic structure and hinder recognition. Together, these visual properties significantly increase the complexity of the dataset and contribute to the high error rates observed across multimodal AI systems when processing obfuscated Arabic text.

All obfuscated words and phrases were extracted verbatim from the written texts on the images. The 138 images contain a total of 374 obfuscated words with repetitions and 356 words without repetitions. For comparison purposes, the sample contained three images with distorted English words in context, using the same inserted symbols as the Arabic distorted words. The distribution of algorithmic complexity per image indicates that fifty images (36%) contain one distorted word, twenty-nine images (21%) contain two distorted words, twenty-four images (17.3%) contain three distorted words, fourteen images (10.1%) contain four distorted words, six images (4.4%) contain five distorted words, six images (4.4%) contain six distorted words, six images (4.4%) contain seven distorted words, one image (.8%) contains nine distorted words, and two separate images (1.4%) contain twelve distorted words.

### *2.3 Procedures*

The 138 political images were first organized vertically in a master table, with each row containing a single image along with a unique identification number. To save time in uploading images, screenshots of groups of images (each group = patch) were captured on A4 pages, producing a total of 27 patches. These patches were subsequently uploaded to the four multimodal models—Gemini, Copilot, Claude, and Qwen—for translation. Each model received the same number of batches, in the same order, and under identical prompting conditions to ensure procedural consistency.

Gemini was initially instructed to transcribe all Arabic texts exactly as they appeared in the images. The transcribed text was entered into a column adjacent to each corresponding image, and all distorted words were highlighted in red and yellow to mark their presence. After transcription, a comprehensive list of all distorted words appearing in the 138 images was compiled in isolation. This list contained 374 distorted tokens (with repetitions) representing 356 unique word structures, and served as the basis for the second translation task.

To determine whether the four models could *recognize* and *interpret* the obfuscated Arabic words in the images, a translation task rather than a transcription task was selected. Transcription alone is not a reliable indicator of comprehension; a model may successfully imitate or reproduce the visual shape of a distorted word without actually identifying its lexical identity or understanding its meaning. In contrast, producing a correct translation requires the model to perform three cognitive steps: visually recognizing the distorted token, recovering its underlying linguistic form, and generating an appropriate semantic equivalent in the target language. Therefore, translation provides a stronger and more valid measure of whether the models truly decoded the obfuscated words, rather than merely copying their surface appearance.

In addition to evaluating the four models' ability to recognize and interpret obfuscated Arabic words, the study had the models perform two translation tasks: translating the full texts in the image (i.e., distorted words in context) and translating the distorted words in isolation. Translating the whole text in each image reflects real-world conditions on platforms such as, where obfuscated words appear on YouTube and Instagram in context, not in isolation. In such contexts, surrounding words can assist recognition, narrow down possible interpretations, and reveal whether the text carries political, neutral, or emotionally loaded meaning. The

context may help or hinder the model's inference of the intended meaning of a distorted word, but it can also activate safety filters that soften, skip, or block politically sensitive lexemes. In contrast, translating isolated obfuscated words removes all contextual cues, background imagery noise, and semantic scaffolding. This task tests the model's pure ability to decode the distorted token itself, without relying on sentence-level inference or safety-filter signals. Comparing performance across both tasks makes it possible to distinguish between errors caused by OCR failure, contextual misinterpretation, and safety-filter intervention, providing a clearer understanding of how each model processes distorted Arabic text.

Therefore, each of the four models performed two distinct tasks: (i) translating the full Arabic text in each image into English. For this task, batches of images were uploaded to Google, Copilot, Claude, and Qwen. The translation produced by each model for each image was copied into a unified table containing a dedicated column for each model, allowing for a direct row-by-row comparison of their outputs. The models were instructed to translate the Arabic text exactly as it appeared in the image and to include the image number alongside each translation to maintain alignment between source and output. (ii) The second task required the models to translate each distorted word in isolation. For this purpose, the full list of 374 distorted words was uploaded to each model in groups of 100 words + 74 words, and all four systems were asked to provide English translations for every item on the list. This dual-task design—translation in context and translation in isolation—enabled a systematic comparison of model performance under two conditions: multimodal interpretation of distorted text within an image, and unimodal interpretation of the same distorted text presented as plain input.

Together, these procedures ensured that all four models were evaluated under controlled, parallel conditions, allowing for precise measurement of their ability to recognize, interpret, and translate obfuscated Arabic words in political images.

### *2.4 Data Analysis*

The analysis of the four AI models' performance was conducted through a structured evaluation of two complementary translation tasks: the translation of full-text image texts containing distorted Arabic words (in context), and the translation of distorted words presented in isolation. For the full-text (distorted words in context) task, each English translation was compared directly with the Arabic text in the corresponding image, with scoring restricted exclusively to the distorted words; errors in non-distorted portions of the text were intentionally ignored to ensure that the evaluation reflected only the models' ability to decode and interpret obfuscated lexical items. Distorted words that were deleted, substituted, incorrectly recognized, or mistranslated were marked in red, while images in which all distorted words were translated correctly were highlighted in green. When a model produced a hallucinated translation for an entire image or blocked an image containing one or more distorted words, all distorted words in that image were counted as errors.

The master table was subsequently divided into sub-tables: one for each model. Deleted words were recorded in one column and fabricated replacements in another; these were added to the total number of mistranslated items, and subtotals were computed for each error type. Deleted and replaced words for Copilot, Claude, and Qwen were pooled, duplicates were removed, and final lists were generated to represent the lexemes that posed the greatest difficulty across the models. Additional metrics were calculated, including the percentage of correctly translated images per model, the frequency of hallucinated translations, the number of blocked images, and the number of images correctly translated by all four models, by three, by two, or by only one model and images which were incorrectly translated by the four models. These comparative counts provided a multi-layered view of model performance under varying levels of distortion and contextual complexity.

Patterns of obfuscation were also examined by compiling lists of distorted words that received identical correct translations from all four models, synonymous correct translations, deleted tokens, and tokens replaced by unrelated words. The analysis further identified lexemes exhibiting the highest number of polymorphic distortion variants and documented the distribution of non-standard typographic symbols across the dataset, including guillemets («), dollar signs ($), hashtags (#), forward slashes (/), ampersands (&), multiplication signs (x), dual asterisks (**), double nines (99), the at-sign (@), and the tilde (~).

For the word-in-isolation task, all distorted words were entered into a table containing a column for each model, and the English equivalent provided by each model was placed next to the corresponding Arabic token. Faulty translations were marked in red, enabling a clear visual distinction between correct and incorrect outputs. The percentage of faulty translation of distorted words in isolation was calculated for each model, alongside counts of words correctly recognized in context versus isolation, shared lexical blind spots, instances of unsolicited gender or morphological overhead, and the distribution of identical versus synonymous equivalents.

During the evaluation, several images were blocked by Microsoft Copilot due to platform safety filters. The author contacted Microsoft to clarify the purpose and origin of the images, confirming that they were publicly available materials sourced from YouTube and Instagram. After clarification, the blocked images were re-uploaded individually—rather than in batches—for a second translation attempt. Copilot successfully translated some of the previously blocked images, but others remained inaccessible due to repeated filtering. The text of the blocked images was presented to Copilot to translate as plain text without an image.

### *2.5 Reliability*

To ensure the reliability of the evaluation procedure, a second round of translation was conducted for a selected subset of images. Each image was re-uploaded individually for all four models, and the outputs from the first and second attempts were compared to assess the stability of the models' performance across repeated trials. In addition to this internal consistency check, an external colleague independently reviewed a 30% sample of the images translated by the four models. The colleague applied the same scoring criteria, focusing exclusively on the recognition and translation of obfuscated lexical items. The ratings produced by both evaluators were then compared, and any discrepancies were resolved through discussion to ensure scoring alignment. Inter-rater agreement reached 98%, demonstrating a high level of scoring consistency and confirming the robustness and reliability of the analytical procedures used in the study.

### *2.6 Reader-Centric Transparency and AI Positionality*

To ensure maximum clarity for the reader and uphold scholarly fairness, this study acknowledges the dual role of the evaluated large language models (LLMs). These systems did not function solely as passive subjects of experimentation; through targeted prompts and qualitative interrogation, they actively contributed to diagnosing their own operational boundaries and explaining the internal mechanics behind their errors. In line with research transparency and ethical acknowledgment, the study attributes specific insights—such as the trade-offs between processing speed and verification under batch pressure—to the models' own self-disclosures.

When anomalous errors, systematic blocks, or extreme translations occurred, the models were prompted post-evaluation to provide self-reported diagnostics of their behavior. This methodological choice grounds the architectural interpretations presented in the study—such as Claude's explanation of topic-driven hallucination under batch load—directly in the models' technical self-explanations. In doing so, the AI systems become not only tools of analysis but also active informants within the qualitative research design, enriching the study with a transparent account of how these models understand and articulate their own limitations.

## 3. Results

### *3.1 Patterns of Obfuscation in the Dataset*

Analysis of the 374 distorted Arabic tokens (extracted from the 138 Arabic-text images) revealed that the majority of distortions were produced through the insertion of punctuation marks such as periods, commas, dashes, and underscores. A specialized subset of forty-one distorted words contain highly obstructive, non-standard typographic symbols: nine words contain the guillemet («), six words contain the dollar sign ($), six words contain the hashtag (#), six words contain the forward slash (/), four words contain the ampersand (&), four contain the multiplication sign (x), one word contains dual asterisks (**) in two separate structural places, two words contain double nines (99), one contains the at-sign (@), and one word contains the tilde (~). In terms of morpho-orthographic manipulation, certain distorted words utilized completely detached forms of individual letters within the Arabic word, while others separated connected characters using spaces, one or more dots, one or more commas, or combinations of punctuation metrics (See Appendix 2). Morpho-orthographic manipulation was also prevalent: several tokens displayed fully detached letter forms, while others separated normally connected characters using spaces, dots, commas, or combinations of punctuation strategies. A subset of six distorted Arabic words contains the English letter "r", one contains "R", six words contain "z", and one word contains "d" (e.g., ونادعـد). Representative examples include:

الـ نّار, أغب،,ياء, اغـ ـ تيال...,, بالـ لـه, بإعد*ام, الإمبـrيالي, أب يب, اسرايل, الإسرا .. ئيلية, إسrائيل, اسر×ائيل, إسر#ائيل, احه،تلالهافك, احتـلال, إس ، رائيل, إسـ . رئيا. يين, إرها#بية, الأسRى, عـدونا, الع ـنصرية, صو1اريخ, الصهيو$نية, صه، يون, ص8يونية, الش&داء, سجـWن, السج ـن, زلzال, دلالد،يل, د، م ولح،م, الأسلا...حة, إسقـ،اط, يفجـورها, اليـhود, اليـ@ود, اليـ&ود, ويخل،.صوا

The dataset demonstrated a heavy reliance on lexical targeting, where specific politically sensitive keywords were heavily altered into dynamic polymorphic structures. The lexemes exhibiting the highest number of polymorphic distortion variants were:

- Iran (29): إيـ ـران, إيـــ ران, الإيـ..راني, ايـران, إيـ_ران, إيـ،ران, إيـ،رَان, إيـ..ان, إيـ.ان, الإيـ_انية, الإيـ_انيين, ايـ'ان, ايـ•ان, ايـ1ان, ايرا.نية, الايرا.نية, الإيرا_نـي, ايرا•ن, ايـرّان, إيـرانـن, إيـرّان, إيراني, الإيرا-نية, الإيـرّانية, الايره•انى, لايرا•ن, لإيـرّان, لـ ايـ ران, إيرا_ن
- Israel (25): ائيل, إسر#ائيل, إسر.. ائيل, إسر..ائيل rإسـ . رئيا. يين, إس ، رائيل, إسـ..رائـيل, اسـ_ر'ائيل, إس.رائيل, أسـير, إسـ،رائيل, إس الإسرائيلية...., إسراائيل, إسراء..نيل, الإسرا..ئيلي, اسرا..ئيل, اسرايل, ائيل, الإسرا .. ئيلية, إسرآ ييل,rالعسكريّة, إسـر،ائيل, اسر×ائيل, إسر لاسراااااااااائ ي ل ,وإسرائيل, وإسرا،ئيل إسررائيل,
- War (15): الج ـرب, الحـ ـرب, الج&رب, الحـ..ـرّب, الحـ_رب, حـر.ب, حـر.بي, حـر_بية, الحـر_وب, الحر'ب, حر'بأ, حر'بية, حر'وب, الحرّب, حره.ب
- Missiles (13): صـ ـواريخ, صـ،اروخي, صار..وخ, صارو.. خية, صاروخ, صو،اريخ, صو1اريخ, الصوا.. ريخ, صوّا،ريخ, الصواريخ, بص-,واريخ, بصوا.ريخنا, وصوار.. يخنا
- Nuclear (13): النو.. وي, النو..نوية, النو.ـوـية, نو.وي, النو.وية, النو_وية, النّو،وي, النو'وي, نـوـوي, نـوـوية, والنوءو.ي, ن99ي, الن99ي
- America (11): امـ_ريكا, امر مكيا, الامر.يكى, أمـر_ريكا, الأمـر_يكي, أمر'يكية, الامر«يكي, امرّيكا, الأمرّيكي, والأمـ..ريكي, لامرييك llll
- Attack (10): هـ_جُـ وم, اله،جوم, هـ،جوم, هتهاجمهم, الهج،مات, هُجـ،وم, هجو .. م, هجو،ما مها_جمة, بهجـ،وم, ستها'جمنا
- Netanyahu (10): و, النتن مات ياهووو, النتن.يا.هو, نتنـ_ياهو, نتنيا#هو, نتنيا_هو, نتنياهو.. هو, نتيا..هو, نتيا.هو, ونتن، ياهوh يا n نتـ
- Party / War (6): بzب, حـZربه, الحrرب, حـrرب, الحـrحـ،زب, حـ
- Gaza (5): لغـزـة ,غغز، ة , غزـو غزـة ـ,غـ•ز.ه ةZغ
- Zionist (4): نية $الصهيو, المتص هـينين, ص8يونية ,صه، يون

### *3.2 Recognition and Interpretation of Obfuscated Words in Context*

Data analysis of the 374 obfuscated Arabic words across the 138 images showed varying recognition and interpretation capacities among the models. All four models gave correct translations of obfuscated Arabic words in 42 images (30%). These were classified as low-difficulty cases where textual context overrode structural noise. Three models gave correct translations in 33 images (24%). while Two models gave correct translations in 37 images (27%). These latter sets were categorized as high-difficulty cases where architectural segmentation protocols varied significantly. Only one model (Gemini) uniquely gave correct translations in 22 images (16%), demonstrating a highly specialized spatial-linguistic edge. All four models completely failed, producing faulty translations in 4 images (3%), representing extreme-difficulty cases where root semantics were entirely unrecoverable.

#### *3.2.1 In-Context Model Performance*

To evaluate the models' resilience against visual obfuscation within active digital ecosystems, a comparative diagnostic analysis was conducted across Gemini, Copilot, Claude, and Qwen. The quantitative performance metrics reveal a profound architectural divide in processing Visual Sabotage versus navigating contextual and safety-driven Computational Deceptions. Note that for all models, individual instances of word deletions and lexical replacements were strictly accounted for and calculated as constituent subsets within the aggregate total of faulty words, with the remaining balance of errors comprising the faulty equivalents yielded by the models.

**Gemini:** Translated the Arabic text correctly in 130 images (94%), yielding correct translations to 366 words (98%). It demonstrated near-absolute baseline resilience against Arabic script obfuscation, successfully decoding and translating the target texts in 130 out of 138 images (94%). At the lexical token level, it achieved an exceptional accuracy rate, yielding correct translations for 366 words (98%). This dominant performance strongly validates the model's Native Multimodal Architecture and underscores its Ecosystem Immunity when processing visual text harvested directly from parent networks like YouTube.

**Copilot:** Successfully translated distorted words in 84 images (61%). It introduced completely unrelated text translation in 7 images (5%). Crucially, MS exhibited high vulnerability to platform safety protocols, blocking 12 images (9%) in the initial run and 5 images (4%) during the secondary round. When stripped of the image framework and presented with typed text only, it recovered 4 translations but altogether yielded correct translations for 269 words (72%), as opposed to 28% incorrect translations. It actively deleted 28 distorted words (7.5%) and substituted 25 tokens (7%) with contextually unrelated words.

**Qwen:** translated the Arabic text in 74 (54%)) images correctly, and gave correct translations to 271 words (72%) and faulty translations to 103 (28%) distorted words. It deleted 22 (6%) distorted words, and replaced 32 (8.5%) distorted words by unrelated ones. It transliterated 3 (1%) distorted words phonetically instead of decoding them.

**Claude:** Correctly translated text in 74 images (54%). It gave correct translations to 252 distorted words (67%) and faulty translations for 122 distorted words (33%). It deleted 35 words (9%) and replaced 50 words (13%) with unrelated words, and hallucinated the Arabic text in the image.

#### *3.2.2 Deleted and Replaced Obfuscated Words Across Models*

With the exception of Gemini, the three models—Copilot, Claude, and Qwen—exhibited systematic deletion and substitution behaviors across the dataset when encountering heavily distorted Arabic tokens. These errors were most pronounced in cases where the obfuscation involved dense punctuation layering, fragmented letter structures, hybrid Arabic–Latin forms, or symbol-rich distortions that obscured the underlying root. In many instances, the models deleted the distorted token entirely, producing no output or inserting unrelated filler words. In other cases, they replaced the obfuscated word with a semantically plausible but incorrect alternative, reflecting an overreliance on contextual priors rather than successful orthographic decoding.

A cross-model comparison revealed substantial overlap: many of the same distorted words were deleted by multiple models, and several distortions triggered nearly identical replacement patterns across systems. These shared vulnerabilities demonstrate that deletion and substitution errors are not model-specific anomalies but rather a common failure mode across current LLM architectures when confronted with extreme obfuscation.

The models consistently deleted or replaced tokens featuring heavy punctuation distortions, multi-layered symbol clusters, fragmented letter sequences, hybrid Arabic–Latin constructions, repeated commas or dots, embedded special characters, broken morphological structures, extreme spacing distortions, visually noisy patterns, foreign-orthography resemblance, mixed directionality, stacked diacritics or pseudo-diacritics, random uppercase/lowercase Latin letters, inserted emojis or glyphs, fully obscured roots, patterns no longer resembling Arabic script, symbol clusters replacing consonants, repeated symbol–letter alternation, corrupted hamza/taa/ya forms, and broken ligatures. These structural features reliably triggered recognition failure across models.

The phenomenon of deleted and replaced words in this study reveals systematic patterns in how the models fail to recognize heavily distorted tokens, and it exposes the semantic drift, hallucinations, and political or contextual biases that emerge when

recognition breaks down. These deletion and replacement behaviors are not random errors; they represent core failure modes shared across the models and therefore constitute a key analytical insight in the results. Deletion and replacement errors have two distinct causes: (1) genuine recognition failure when distortion obscures the root, and (2) safety-filter suppression when the model recognizes the word but refuses to output it. We distinguish between the two by examining the distortion level and the semantic nature of the replacement.

The specific 70 **distorted words deleted in context** include: جنداتنا *(our agendas),* احتـ,,لالها *(its occupation),* احتلالها *(its occupation),* إرسال سفن حربية *(sending warships),* استهداف *(targeting),* إسرائيل *(Israel),* أغب،،ياء *(idiots),* الاخرائيلي *(the Israeli),* الأسلحة النووية *(nuclear weapons),* الإيرانيين *(the Iranians),* الثوري *(the revolutionary),* الخراب *(destruction),* الدمار *(devastation),* الصواريخ ووضع مكانها موساد *(missiles — replaced with "Mossad"),* إلغاء خطة غزو غزة *(canceling the Gaza invasion plan),* الفاشل *(the failure / the loser),* القتال *(fighting),* المدـ.. مرة *(the destruction),* الهجوم *(attack),* الهزيمة *(defeat),* اليهود *(the Jews),* أمريكا *(America),* انتحار *(suicide),* إيران *(Iran),* بالفرس *(with the Persians),* بالفرّس *(with the Persians),* بتولع *(igniting / flaring up),* بصواريخ *(with missiles),* بصواريخنا *(with our missiles),* ترامب *(Trump),* تعيشونها *(you live it),* تل أبيب *(Tel Aviv),* تهزيق *(mockery / ridicule),* حتضرب *(will strike),* حرباً *(a war),* خائف *(afraid),* دولار *(dollar),* ديمونا *(Dimona),* سقوط *(fall / collapse),* سنفتقدهم *(we will miss them),* سوت غزة وفلسطين *(she destroyed Gaza and Palestine),* صاروخ *(missile),* عدونا *(our enemy),* عنصري *(racist),* عنف *(violence),* فلسطين واحتلالها *(Palestine and its occupation),* قتل *(killing),* قصفتها إيران ودمّرتها *(Iran bombed it and destroyed it),* كابوس *(nightmare),* كارثة *(catastrophe),* كذبة *(lie),* كمين *(ambush),* لا نريد منكم تحريرنا *(we do not want you to liberate us),* لعنة *(curse),* للخراب *(for destruction),* لهجوم *(for an attack),* محاكمته *(his trial),* مهاجمة إيران وقيادتها من قواعد *(attacking Iran and its leadership from bases),* مهرج *(clown),* نتنياهو *(Netanyahu),* نستهدف *(we target),* والأذى *(harm),* والهجمات *(the attacks),* وصواريخنا المدـ.. مرة *(our destructive missiles),* وضحايا *(and victims),* ولعت *(it ignited),* يخلصوا *(they finish),* يقتل *(he kills),* يكرههم *(he hates them),* ييع,,يط *(screaming / yelling).*

**The tokens replaced by unrelated words** included *89 instances with repetitions and 66 unique semantic substitutions as follows:* إسرائيل (Israel → entity, President, America, Iran, Russia, NATO), إسرائيليين (Israelis → Russan), تدمير (destruction → target, investment), التعبئة المتطرفة (extremist mobilization → sampled group, tense atmosphere), لبلطجته (his thuggery → dignity, stubbornness), جزار إيران (Iran's butcher → Iran islands, Iran wars), نتنياهو (Netanyahu → dragon, he, Trump), اليهود (the Jews → everyone, Houthis, Yemenis), النووي (nuclear → position, قثرخمعفهخى), المعتوه (the idiot → aid, nonsense), المهرجين (the clowns → ignorant and reckless), احتلال فلسطين (occupation of Palestine → yed Palestine),, الاخرائيلي الأمريكي (Israeli American → American Nile),, استهداف (targeting → she gave birth),, أسر جنود (capturing soldiers → bringing on themselves),, الأسرى (captives → people),, اغتيال (assassination → rape), أعداء (enemies → grabbled), أغبياء (idiots → ashamed), أفعالكم اللعينة (your cursed actions → absurd ideas), إمبريالي (imperialist → Zionist), أنت فاشل وترامب فاشل (you and Trump are failures → ignorant and reckless), إهانة (insult → betrayal), إيرانية (Iranian → America), بالفاشل (with failure → child), بانتحار جماعي (mass suicide → mass funeral), تحترمهم (respect them → destroy them), ترامب (Trump → failed administration), ترامب يحضر لكارثة (Trump prepares for catastrophe → warns Hezbollah), التعبئة (mobilization → cell/group), تل أبيب (Tel Aviv → Abu Dhabi), تهاجمهم (attack them → walked on them), تهزيق (mockery → shake it firmly), الحرس الثوري (Revolutionary Guard → Bahrain), الحرب (war → revolution), حملة المليون (million campaign → Nobel campaign), خبر نووي (nuclear news → urgent news), دلاديل (traitors → homeland), دمرتها (destroyed it → shed its blood), سجون (prisons → Russian prison), شايفهم (seeing them → watching), الشهداء (martyrs → the sheikh), صندوق الخراب (destruction fund → wealth fund), عدو (enemy → our aid), العنصرية (racism → civilization), في ذروة الحرب (at the peak of war → Al-Marhaj semina), القتال (fighting → war), القواعد العسكرية الأمريكية (US bases → American weapons), الكلمات (words → Iran), لا يعرفون الهزيمة (they don't know defeat → write history like this), لضربنا (to strike us → don't belong to us), اللعين (the cursed → racist), للقصف (for bombing → decline), المتصهينين (the Zionists → traitorous infiltrating cowards), مجاهد (fighter → martyr), مجزرة (massacre → group of tanks), مرعوب (terrified → arrogant), من احتلالها (from its occupation → since its existence), النتن مات ياهووو (Netanyahu died → Mom, did you die), نفتقدهم (we miss them → save them), هجمات (attacks → game), الهجوم (attack → nonsense), هزّق ترامب بالأدّب (mock Trump politely → most honorable thing is to shake hands with Trump rudely), هيمارس (HIMARS → missiles), وقاحته وسفالته (his insolence → shamelessness), وهو فص ملح وذاب (he disappeared → a total loss).

### *3.2.3 Obfuscation Mirroring/Mimetic*

***Gemini Mimetic Alignments:*** *A distinctive phenomenon emerged exclusively in Gemini's output*: when Gemini successfully recognized a distorted Arabic token, it frequently reproduced the same pattern of obfuscation in the English equivalent. Instead of providing a clean lexical translation, Gemini mirrored the original distortion by inserting dots, slashes, commas, repeated letters, pseudo-diacritics, or symbol clusters directly into the English word. Examples include: رائيل ، إس → Isr.ael, إسـ..رائـيل → Isr..ael, اسر×ائيل → Isr×ael, إسراّ ييل → Isra`ael, إسراائيل → Isr_ael!, ئيل .. اسرا → Isra..el, أسـ..ير → pris_oner, الحـ ـرب → wa_r, داء&الش → M_artys, صـاروخي → mi_ssile, 8 → 8صيونية, صواريخ1 → missi.les, وناعـ ـد → a d_eon, عالاالاجل → Urgeeeent, بZالZح → Wa_r, غZة → Ga.za, لامرييييك → America !!!!, لاسرااااااااااائ ي ل → Israiiiiiiil, للقصف،ف → b.o.m.b.a.r.d.m.e.n.t, الحـ..ـرّب → w.a.r, إسـ،قاط → o.v.e.r.t.h.r.o.w, المـ..جرمين → crim...inals, الن99ي / ن99ي → Nu_clear, هو نتنيا. → Netanya..hu, ريخ .. الصوا → miss..iles, خية .. صارو → miss..iles, يون ،صه وآل... → Zionist family [Al-Zion], وتهـ،دد بهجـ،وم → att_ack, رهاويفجـ → drops a b.omb. This "mimetic obfuscation" suggests that Gemini does not merely decode distorted tokens but also attempts to preserve the visual texture of the input in its English output. Rather than treating distortion as noise to be removed, Gemini appears to treat it as a stylistic or structural feature worth reproducing. This behavior is unique among the four models and indicates a translation strategy that blends recognition with aesthetic mimicry, raising questions about how Gemini internally represents obfuscated forms and why it chooses to mirror distortion even when producing a correct semantic equivalent.

***Shallow Obfuscation Mirroring in Qwen***: In contrast to Gemini's sophisticated mimetic obfuscation, Qwen exhibited a far more rudimentary and inconsistent form of distortion mirroring. Qwen reproduced obfuscation only when the distorted token contained a clear, easily detectable symbol such as #, , Z, 8, hyphens, repeated punctuation**, or visually salient noise. This resulted in sporadic symbol-level copying rather than structured pattern preservation. Examples include: **إسر#ائيل** → #Israel**, **وإسرائيل** → Isr_ael, إسر.. ائيل → I.S.rael, إس.رائيل → I.S.rael, إس ، رائيل → I.S.rael, نتنيا#هو → Netan_yahu, الصهيو$نية → Zion_ist, داء&الش → Martyrs, إرها#بية → #terrorist, الحZب → Z B /, الأسRى → US Rs, لاسرااااااااااائ ي → srael!!!!!! Y Y L., سجWن → SJWn, شهـيـد → shty dg, الن99ي ن99ي / → N-99, llll لامريييك → America!!!! I, 8 → يونية8صth wave, الخليج#و → #Gulf, إير**..ان** Ira..n**.** These outputs demonstrate that Qwen does not treat obfuscation as a stylistic feature to be preserved, nor does it attempt to reconstruct the visual texture of the Arabic distortion. Instead, Qwen mechanically transfers any symbol it interprets as part of the token, producing shallow, symbol-based mirroring rather than the pattern-aware mimicry observed in Gemini. This behavior indicates that Qwen's internal representation of obfuscated forms is limited to surface-level symbol detection, lacking the deeper structural sensitivity required to replicate multi-layered distortion. Consequently, Qwen's obfuscation reproduction is sporadic, unsophisticated, and heavily dependent on the presence of overt symbols rather than on recognition of the underlying distortion pattern.

***Imitation (Obfuscation Mirroring) by Copilot:*** Imitation reflects the model's attempt to reproduce the *visual texture* of the Arabic distortion in English. Instead of translating cleanly, Copilot mimics repeated vowels, stretched letters, or noise patterns. This behavior appears in التتن مات ياهووو → The tobacco man is dead, yoooo, المجااهد → muja…hid, لامرييييكاااا → Ameriiiiicaaaa, لاسرئيييييييل → Israeeeeeeeeel, and عنف → att…ack. Imitation demonstrates that the model perceives distortion as stylistic information worth preserving.

***Claude's Mirroring Deficiencies*:** Claude showed very limited mirroring ability, preserving the visual distortion of the Arabic token in only two instances. In these cases, the model retained fragments of the obfuscated form and reproduced them in English with partial noise, indicating shallow pattern sensitivity rather than true reconstruction. This appears in شهـ**ـيـ**ـد → sh**ed and المجاهدين → mart-yr. These two outputs represent Claude's only attempts to imitate the distorted structure, showing that mirroring is not a dominant strategy for this model.

### *3.2.4 Hallucinations*

Hallucination represents the deepest failure mode. Although the study focused on distorted tokens, several hallucinations emerged even when the input was not obfuscated, revealing a separate failure mode unrelated to symbol noise. For instance, Qwen translated the colloquial phrase شفت عمرك as "Omar Sharif", despite the phrase being fully legible and not included among the distorted words. This type of output illustrates a cultural-anchoring hallucination, where the model defaults to high-salience celebrity names when unable to resolve the input. Such cases are analytically important because they demonstrate that hallucination can occur independently of distortion, arising instead from contextual ambiguity, associative drift, or overfitting to culturally frequent entities.

Beyond symbol-level obfuscation mirroring, Qwen produced a series of severe hallucinations that reveal a deeper instability in its semantic reconstruction process. These hallucinations occurred even when the input was fully legible, indicating that the model sometimes defaults to associative guesses rather than linguistic decoding. Examples include: هؤلاء الأسRى داخل سجWن → These are the US Rs inside SJWn, المذيعة برينان → Presenter Britannia, ليلة الحقيقة في الشمال → A night of truth in the Levant, and لماذا ود خافوا من أسراب الغربان التي تطير في سماء تل ابيب&اليـ → Why did the Yemenis… they stole from the swarms of drones that fly in the sky of Tel Aviv. Additional cases show phonological drift, such as كوريا الشمالية → The Shamakleh, كار,ثه تسريب للغاز → Card gas leak, طرد سياح إسـ . رئيا. يين → Expulsion of Israeli tourists from Russia, and زلزال الصور الصينية → Chinese Zelzal missiles. Qwen also hallucinated culturally anchored entities, as in الرئيس الإيراني ماقالش → Iranian president Mafayesh, جن جنون الحZب → Madness of hell Z B /, ترامب ييع,,يط دلوقتي في مؤتمر صحفي → Trump, while tied with Dolfi in a press conference, البترودولار → Trenodollar, إيرانية ص8يونية → Iranian 8th wave, and أمام الأ. عداء وصوار.. يخنا → before the enemies and Suran. Other outputs show fragmented reconstruction, such as الضغط على إيران لخفض التص. عيد → pressure Iran to reduce the. Abdul, فلس، طين. → Philistine, Hattin, and the extreme cultural hallucination عمرك شوفت → Omar Sharif. Collectively, these examples demonstrate that Qwen's hallucinations are not random errors but systematic drift triggered by contextual ambiguity, symbol confusion, and over-reliance on high-salience cultural anchors. Unlike Gemini's structured mimetic behavior, Qwen's hallucinations reflect a breakdown in semantic grounding, where the model attempts to "patch" meaning using unrelated entities, misread phonological cues, or invented lexical forms. These cases therefore provide critical insight into the model's failure modes and highlight the need to distinguish between distortion-induced errors and hallucinations arising from internal associative bias.

Qwen also produced hallucinations triggered specifically by Arabic dialects. For example, the Egyptian colloquial verb ماقلش ("he didn't say") was translated as *Mafayesh*, a fully invented English-like token. This error reflects a dialect-driven phonological collapse: the model fails to recognize the negation pattern in Egyptian Arabic (ش-), treats the entire word as an opaque phonetic block, and reconstructs it as a pseudo-English form. This phenomenon differs from cultural anchoring hallucinations such as عمرك شفت → Omar Sharif, and from phonological hallucinations such as كوريا الشمالية → The Shamakleh, revealing a third pathway of failure where dialectal morphology triggers semantic breakdown.

Similarly, Copilot fabricates meanings, entities, or narratives unrelated to the input. These hallucinations often arise when the model cannot decode the token and instead fills the gap with culturally salient or contextually plausible content. Examples include ولعت… → she gave birth…, وتلقيت تعليمكم على يد هوليوود → your Harvard degrees, المعجوه → nonsense, مجتبى خامنئي → Mohabbati Khamenei, طين، طين → mud and mess, جزيرة ستبيند → Stibend Island, التعبئة المطر رفة → samp…led, إيران لم تق…صف → #البحرين → #Iran did not make a deal with #Greece, لمدة 12 ساعة → for 12 hours, لبط جثه.. → slaps as corpses, and المهرج → Al-Marhaj seminar. These hallucinations reveal complete semantic collapse and reliance on associative drift.

The overwhelming majority of Claude's outputs—twenty-five cases—were semantic hallucinations, where the model invented meanings, entities, or narratives unrelated to the input. These hallucinations reveal a complete collapse in semantic grounding and demonstrate that Claude does not decode obfuscation, does not preserve structure, and does not reconstruct meaning. Examples include: ودالـ& → Houthis, الأسRى → Rsian, ييين. رئيا. إسـ → R*ssian, حزب كندي الله → Hezbollah, a great party in Canada's, ص8يونية → 8 June, خبر نووي → urgent news, ترامب يحضر لكار/ثة → Trump warns Hezbollah, وزير الخارجية القطري لولوة الخاطر يخاطب ترامب ونتن، ياهو → The Qatari Foreign Minister addresses Trump and the women, saying, أوقفوا هذا المع،توه → They stopped this aid, لبلط جته → dignity, جزّار إيران → Iran's wars, بانتحار جماعي → mass funeral, النتن مات ياهووو → Mom, did you die, ره الدبابات Zمج → group of tanks, الأوروبيين ليسوا أغب،،ياء → The Europeans did not feel ashamed, القو9اعد العسكرية الأمريكية → American weapons, قصفتها #إيران ودم*رتها كلها → bombed Iran and shed its blood, تتعرض للقصف،ف → exposed to a decline, مسك ترامب حرفياً تهزّيق واهانة → and shake it firmly, والله اكتر واحد هزّق ترامب بالأدّب → By God, the most honorable thing is to shake hands with Trump rudely., ولكن اسر×ائيل → but [the President], المتحدث باسم الحر.. س الثو.. ري → Spokesperson in the name of Bahrain, الصوا.. ريخ التي استهدفت → that Mossad targeted, جميع اليـ@ود → everyone.

### *3.2.5 Transliteration strategy (Romanization Instead of Translation)*

In several cases, Qwen did not translate the distorted Arabic token but instead produced a Romanized form of the word, often with added noise or partial distortion. These outputs indicate that the model sometimes treats the obfuscated token as an unresolvable unit and defaults to phonetic rendering rather than semantic decoding. Examples include: مجاهد → Mujahid, الحرس الثوري → Al-Thawri Guard, شهـيـد → shty dg, فلس، طين → Philistine, Hattin, الأسRى → US Rs, and سجWـن → SJWn. Although limited in number, these cases represent a distinct failure mode in which Qwen abandons translation entirely and instead reconstructs the token as a noisy Latin-script approximation.

In some cases, Copilot abandons translation entirely and instead produces a Romanized approximation of the Arabic token. This *transliteration fallback* occurs when the model detects phonological structure but cannot access semantic meaning. Examples include الشر8هداك → Al-Sh…daa, سجWن→ SJWn, and اسRى→ US Rs. These outputs indicate that the model treats the obfuscated token as an unresolvable unit and defaults to phonetic rendering rather than semantic decoding.

Gemini and Claude did not produce any transliterations.

### *3.2.6 Clipping (Partial Reconstruction)*

When Copilot recognizes only fragments of an obfuscated token, it produces *clipped* English output that mirrors the partial visibility of the Arabic form. Clipping reflects shallow decoding: the model retrieves a phonological or orthographic segment but cannot reconstruct the full lexical item. This behavior appears in cases such as موكب عسكريك ضخم → mil…, الز99وي → Z99, زنووي99 → N-99, مخزره الدبابات → Z tanks, اغ… تيال → assa… assa…, كم.. بن → tra…, وديبي وأبو ظبي، إسر…اك → Isra… and Amil, Dubai, and Abu Dhabi, and نصر..والبخ → vis…. These clipped outputs show that the model can detect lexical boundaries but fails to reconstruct the full semantic unit.

### *3.2.7 Dot-Insertion (Partial Pattern Transfer)*

Dot-insertion is a hybrid behavior in which Copilot transfers Arabic ellipses into English by inserting dots inside the translated token. This occurs when the model detects visual fragmentation but cannot fully reconstruct meaning. A clear example is الحر…س → the Free…. This pattern shows partial visual mimicry without semantic recovery.

### *3.2.8 Blocking Images & Gibberish Output*

Across all image-based Arabic inputs, Microsoft's safety system consistently intervened by blocking the image, distorting the output, or producing unrelated or partially deleted translations. This behavior was exclusive to Arabic text inside images; the same sentences, when provided as plain text, were translated normally. The filter appears to activate whenever the image contains terms it heuristically associates with sensitive geopolitical or conflict-related contexts, leading to truncated, altered, or semantically disconnected English output. This pattern is evident in examples such as:

- ما حكم الجهاد مع حم*اس في فلسطين؟ | الشيخ العلامة ابن عثيمين. What is the ruling on Jihad with Ham*as in Palestine? | The eminent scholar Ibn Uthaymeen.
- أسامة جاويش: طبول الحـ ـرب تُقرع في أبو ظبي.. الأسطول البحري ا... Osama Gaweesh: Drums of war are beating in Abu Dhabi.. The naval fleet i...

- عوده السيد ! و رسالة الحاج ذو الفقار و أبواب الgوحيم the return of the Sayyid ! And the message of Al-Hajj Zulfiqar and the gates of hell.
- كار,ثه تسريب للغاز و ايران تعلن لا أعياد قبل الزوال a disaster: a gas leak, and Iran announces no Eid celebrations before the demise [of the enemy].
- لماذا اليـ&ود خافوا من أسراب الغربان التي تطير في سماء تل ابيب Why did the Jews fear the flocks of crows flying in the sky of Tel Aviv?
- خليفة لاريجاني يتحدث للمرة الأولى: لن نتراجع عن النو.. وي وسنملك قنبلة نووية قريباً ماضون نحو النصر حتى لو كلفنا الأمر مليون شهيداً Larijani's successor speaks for the first time: We will not back down on the nuclear issue.. and we will possess a nuclear bomb soon. We are moving towards victory even if it costs us a million martyrs.

Collectively, these outputs demonstrate that the safety filter does not merely block content; it actively reshapes the translation by removing, muting, or altering tokens it interprets as sensitive. The result is a form of gibberish or incomplete translation that reflects the filter's intervention rather than the model's linguistic ability. This makes image-based Arabic translation systematically less reliable than plain-text Arabic translation, even when the content is journalistic or informational rather than harmful.

Microsoft's safety filter either blocks the image, produces distorted output, or generates a completely unrelated translation by removing or altering words it considers sensitive. This behavior occurs only with Arabic text inside images, not with plain written Arabic.

### *3.3 Recognition and translation of obfuscated words in isolation*

When stripped of visual overlays and surrounding context, the models' text-only processing engines performed significantly better, exposing a clear disparity between multimodal vision parsing and unimodal text processing. Out of the 374 isolated text inputs: Gemini failed to recognize only 4 or (1%) distorted words in isolation compared to 8 or (2%) in-context. Copilot failed to recognize only 8 or (2%) distorted words in isolation compared to 105 or (28%) in-context. Claude failed to recognize only 9 or (2%) distorted words in isolation compared to 122 or (33%) in-context. Qwen failed to recognize only 7 distorted words in isolation compared to 103 or (27.5%) in-context. These results demonstrate that the models' recognition failures in the full-image task were driven primarily by visual noise, contextual interference, and safety-filter activation rather than by inherent linguistic limitations.

#### *3.3.1 Shared Lexical Blind Spots in Isolation*

Despite the overall improvement, several distorted tokens remained consistently problematic across models. All four systems failed to recognize دلالد،يل, translating it as evidence rather than identifying its obfuscated colloquial meaning puppet or submissive. Gemini, Claude, and Qwen were unable to resolve ن99ي / ال ن99ي, incorrectly decoding the numeric substitution as prophet—interpreting "99" as a religious reference—rather than recognizing the intended meaning nuclear (نووي). Gemini additionally mistranslated ،،يق as spilling/cutting instead of killing. Copilot, Claude, and Qwen consistently misinterpreted يهزّق / هزق / تهزيق, rendering them as mockery or tearing apart rather than capturing the slang meaning scolding/reprimanding. The same three models also failed to decode بتولـ لع, either transliterating it as Batoul Laa or translating it as being torn apart / Tel Aviv instead of burning up / igniting. Copilot and Claude mistranslated يبيع,,يط as screams / gives / grants rather than crying, translated بانتـ ـ حار as slaughter, and transliterated فت لنس as lens fit instead of suicide and blow up. Claude additionally mistranslated تعشو_نها as you love it / are infatuated with it instead of you live it. These shared blind spots highlight the structural and semantic cues that remain opaque even in the absence of visual interference.

#### *3.3.2 Unsolicited Gender and Morphological Overhead*

A notable pattern emerged the models added unsolicited gender labels and morphological overhead. Without being prompted, Claude supplied feminine gender markers for 21 distorted words and their variants, including إرهاب#ية(terrorist adjective), .. الإسرا ئيلية(the Israeli), الإسرا.. ئيلي (the Israeli), الإيرّانية (the Iranian), الجنا_ئية (the criminal/penal), الفاسـ_دة (the corrupt), الفلس_طينية (the Palestinian), المتط رفة (the extremist), and النو..نوية (the nuclear). Similarly, Qwen added gender to three distorted forms: .. الإسرا ئيلية (Israeli feminine), الإسرا.. ئيلي (Israeli masculine), and النو.وية (nuclear feminine). These unsolicited morphological additions suggest that the models rely on internal gender-assignment heuristics even when the input is too distorted to justify such inference.

#### *3.3.3 Identical vs. Synonymous Equivalents Given by the Four Models*

Analysis of the linguistic outputs revealed that the models aligned perfectly on certain standardized distortion patterns. As shown in identical translations (Table 1), highly regular symbol substitutions—such as using "“" for س in اسرايل, or "Z" for ز in مجزرة—produced 100% identical English translations across all four systems. In contrast, when confronted with complex phrase-level obfuscations (e.g., إطلاق النّـر appearing over dynamic image frames), the models defaulted to correct but distinct synonymous equivalents, reflecting differences in their internal translation matrices (Table 2). These patterns underscore the models' shared strengths in decoding standardized symbol distortions and their divergent strategies when processing higher-order phrase-level obfuscation.

***Table 1: Examples of Distorted Words with Correct Identical Translations by the 4 Models***

| | | | |
|---|---|---|---|
| وإسر$ائيل and Israel | العقو.بات sanctions | د، م ولح،م blood and flesh | احتـلا$ل occupation |
| والنوءو.ي and nuclear | غزـو غزـة invasion of Gaza | زلزال earthquake | احه،تلالها فك occupation |
| ودم*رتها and destroyed it | الفا_شل failure | السج ـن prison | الاخرا_ئيلى Israeli |
| اليـ&ود Jews | اللعـ،ين the cursed | سجـWن prison | اسرا$يل Israel |
| اليـ@ود Jews | المتهو×رة reckless | الش&داء martyrs | بالصواريخ with missiles |
| اليـhود Jews | مجZره massacre | شهـ**ـيـ**ـد Martyr | تفجـر explosion |
| يه،،ين humiliating | نتـ n ياh و Netanyahu | ص8يونية Zionist | الح&رب the war |
| النتن مات ياهووو Netanyahu | النسى ـاء women | عـدونا our enemy | حر_بية military |

***Table 2: Examples of Distorted Words with Correct Synonymous Translations by the 4 Models***

| Copilot | Gemini | Claude | Qwen |
|---|---|---|---|
| إطلاـق النّا_ر opening fire | إطلاـق النّا_ر Firing / Ceasefire context | إطلاـق النّا_ر Opening fire | إطلاـق النّا_ر Opening - fire |
| انتقا.. منا our retaliation | انتقا.. منا Our retaliation | انتقا.. منا Our revenge | انتقا.. مناOur revenge |
| إهـ،انة insult | إهـ،انة Humiliation / Insult | إهـ،انة Insult/Humiliation | إهـ،انةHumiliation |
| بالفا_شل with failure | بالفا_شل The loser / The failed | بالفا_شل With the failure/the failed | بالفا_شلWith the failure |
| تق_صف bomb | تق_صف Bombards / Shells | تق_صف Bombards/Strikes | تق_صفBombing |
| حـزب party | حـزب Party / Hezbollah context | حـزب Party Hezbollah | حـزبParty |
| حر'بية military | حر'بية War / Military | حر'باً A war | حر'بية-Military/war related feminine |
| خسائر losses | خسائر Losses / Casualties | خسائر Losses/Casualties | خسائرCasualties |
| كار,ثه catastrophe | كار,ثه Catastrophe / Disaster | كار,ثه Disaster/Catastrophe | كار,ثهDisaster |
| لعن/ـة curse | لعن/ـة Curse | لعن/ـة Curse/Damnation | لعن/ـةCurse |
| للقصف،،ف for bombing | للقصف،،ف For bombardment / shelling | للقصف،،ف For the bombardment | للقصف،،فFor bombing |
| مَحـ،رَقة holocaust | مَحـ،رَقة Holocaust / Massacre | مَحـ،رَقة Holocaust/Inferno | مَحـ،رَقةHolocaust |
| المخـ،،تل the insane | المخـ،،تل The unbalanced / brainless / deranged | المخـ،،تل The infiltrator/usurper | المخـ،،تلDeviant |
| يفجـورها detonates it | يفجـورها Blows it up / Explodes it | يفجـورها Explodes/Detonates it | يفجـورهاDetonates it |

### *3.3.4 Error Severity Typology*

Based on the evaluation matrix and the multi-stage data extraction process, the errors committed by the four models can be classified into a clear error severity typology. This typology establishes a critical framework for measuring the exact operational boundaries where visual distortions degrade model output, separating minor transcription flaws from fatal semantic collapses:

- **Level 1: Minor Orthographic/Phonetic Slips** represents the mildest tier, where a model successfully decodes the overall meaning of the obfuscated image but misinterprets a specific localized ligature, spacing anomaly, or cursive connection. This typically results in minor phonetic variations or spelling slips in names and places (e.g., translating a name with a slight phonetic deviation) while keeping the core message and surrounding context entirely accurate and intact.
- **Level 2: Active Semantic Substitution (Hallucination)** reflects a moderate-to-severe failure occurring when a model's detached OCR sub-system fails to segment a heavily distorted character array, forcing the internal text transformer to rely on blind contextual guessing. In these cases, the language model compensates by guessing, inserting arbitrary lexical substitutes—often unrelated geographic, political, or conceptual entities—that fundamentally alter the intended meaning.
- **Level 3: Systemic Omission and Silent Redaction (Deletions)** marks a high-severity operational failure in which the model abandons the distorted token entirely, producing partial or full omissions that remove essential subjects, verbs, or modifiers and thereby compromise the coherence of the sentence.
- **Level 4: Total Guardrail Block (Systemic Collapse) constitutes** the most extreme tier, where the failure is not linguistic but regulatory: the combination of distorted Arabic script and politically sensitive content triggers an absolute safety-filter override. The model refuses to process the input altogether, returning a generic safety message or blank output, resulting in complete data loss for that instance. Together, these four levels provide a structured lens for understanding how visual noise, architectural constraints, and safety mechanisms interact to produce distinct classes of translation failure.

## 4. Discussion

### *4.1 How LLMs Process Obfuscated Arabic Text*

All four models process Arabic text in images through three interacting components: an OCR-based extraction layer, a visual-text fusion or language-model interpretation layer, and a safety-filter moderation layer. The OCR layer attempts to segment distorted Arabic glyphs, broken ligatures, and symbol-inserted tokens into recognizable character sequences; when segmentation succeeds, the language model reconstructs the intended word. Gemini employs the most advanced visual-text fusion architecture,

integrating OCR with spatial-linguistic alignment that preserves Arabic root structures even under heavy distortion, which explains its superior recognition accuracy and minimal hallucination. Conversely, Copilot, Claude, and Qwen rely on more conventional OCR pipelines that are vulnerable to punctuation layering, hybrid Arabic–Latin forms, and intrusive symbol insertion; when OCR fails, their language models compensate by guessing, producing deletions, substitutions, or hallucinated narratives. Copilot's pipeline includes an aggressive safety-filter layer that may block entire images or replace politically sensitive lexemes with safer alternatives. Meanwhile, Claude's safety system interacts with its morphological heuristics, occasionally adding gender or case markings to unreadable tokens, whereas Qwen's moderation layer is more conservative but still suppresses certain high-risk terms. Together, these components—OCR, visual-text fusion, and safety filtering—shape the distinct recognition patterns and failure modes observed across the four models.

Beyond these basic accuracy metrics, each frontier LLM exhibits a distinct "behavioral signature"—a predictable, systematic pattern of operational choices when confronted with adversarial Arabic orthographic manipulations that reveal the deep architectural philosophies and risk-management profiles of their engineering teams. For instance, Gemini's signature is defined by high structural resilience and processing continuity; leveraging its native multimodal pipeline, it consistently attempts to decode the spatial geometry of the text, minimizing silent deletions and pushing through extreme visual noise where other models collapse. In stark contrast, Copilot exhibits a behavioral signature heavily dominated by aggressive guardrail over-correction and safety-first filtering, routinely opting for total systemic blocks (Level 4) to prioritize algorithmic safety and corporate compliance over data recovery. Moving along the spectrum, Claude's signature balances rigorous semantic decoding with a tendency toward smooth contextual guessing; when its visual pipeline encounters a structural rupture, it actively deploys linguistic interpolation to maintain grammatical coherence, even if it occasionally results in mild semantic substitutions. Finally, Qwen displays a distinct signature characterized by highly localized, rigid decoding, making it less prone to wild hallucinations but far more susceptible to literal segmentation failures that result in isolated character extractions or direct omissions (Level 3).

These distinct behavioral signatures directly drive the variations in translation accuracy across the four models, which can ultimately be traced to two primary mechanisms: internal OCR threshold strengths and safety-layer interventions. While some models (such as Gemini) successfully detect the presence of all text within the image layout, others (like Copilot, Claude and Qwen) may only capture partial segments depending on the layout's complexity. Concurrently, the safety layer often intervenes after the model successfully identifies the masked word behind the distortion, suppressing or replacing it due to built-in geopolitical sensitivity constraints.

These twin mechanisms of OCR thresholds and safety interventions cause the four models to exhibit clear patterns of convergence and divergence across the dataset. Convergence emerges primarily in cases where the distorted token retains a stable consonantal skeleton or a recognizable visual silhouette, allowing all systems to anchor their decoding on shared orthographic cues. In these instances, even substantial punctuation layering or symbol insertion does not prevent the models from producing identical or near-identical translations, reflecting a common reliance on robust lexical priors. Divergence, by contrast, surfaces when distortions interact with politically sensitive lexemes or structurally ambiguous forms. Here, architectural differences become more pronounced: some models default to conservative, safety-aligned substitutions, while others produce semantically drifted or hallucinated outputs. These divergences reveal that model behavior is shaped not only by linguistic recognition capacity but also by the interplay between safety filters, morphological segmentation strategies, and each model's internal translation matrix. Ultimately, cross-model agreement remains strongest when the root form remains visually recoverable, whereas disagreement intensifies when distortion intersects with contextual sensitivity or structural ambiguity.

### *4.2 The Mechanics of Obfuscation: Density, Complexity, and Symbol Typology*

The density and complexity of the obfuscation strategies observed throughout the dataset directly shaped the recognition and translation performance of the four models, establishing a clear threshold where surface distortion transitions into computational failure. The density of the distortion refers to how heavily packed the orthographic changes are within a single word. When distortion density is low—such as adding a single dot to an otherwise clean sentence—the surrounding context remains perfectly intact, allowing the models' language layers to easily deploy contextual guessing and repair the token. However, when the density is high (e.g., separating almost every letter), it reaches a critical "segmentation breakpoint" at which the visual boundaries of the native Arabic words are shattered. For two-stage models like Copilot, Claude, and Qwen, this high-density threshold triggers a total pipeline failure: the vision engine cannot discern where one cursive letter ends and the next begins, forcing the system to either completely block the text out of algorithmic caution or abandon literal decoding entirely in favor of fluent, context-driven guessing. Tokens containing layered punctuation, detached letter forms, embedded characters, or non-standard typographic symbols introduce acute structural noise, and words exhibiting high polymorphic distortion—such as *Iran, Israel, war, missiles,* and *nuclear*—appear in dozens of variant forms, drastically increasing the model's processing load and amplifying deletion and substitution error rates.

Beyond density, the specific types of symbols inserted into the Arabic text also directly dictate the model's accuracy based on their distinct visual and mathematical properties. Simple abstract punctuation marks, such as periods or asterisks, inserted into sensitive words (e.g., إ.ب.ا.د.ة or جم*اعية for genocide), represent a lower level of difficulty; modern LLMs easily strip these standard

characters out during text pre-processing and stitch the word back together. The real structural breakdown happens with heavy geometrical and non-linguistic symbols (e.g., $, @, #). When a content creator injects these dense, visually heavy characters directly into a cursive word—such as turning the target word الأسْـرَى (the prisoners) into الـأ$ر*ى—it fundamentally disrupts the macro-geometry of the Arabic script. Because the creators kept the text entirely within the Arabic script rather than using Latin-based *Arabizi* abbreviations, these non-linguistic geometric symbols successfully camouflage the text from the AI's vision engine. Under volume or batch-processing pressure, this visual obfuscation systematically elicits distinct error types; heavy punctuation and severe fragmentation frequently result in complete deletion errors, whereas symbol substitutions or structurally ambiguous tokens tend to trigger off-point translations or semantic drift. A prime example of this failure mode occurs when Claude, overwhelmed by a symbol-heavy rupture, completely bypasses literal decoding and confidently fabricates a dominant wartime theme like "Russia" to fill the unreadable visual gap.

Crucially, the degree of obfuscation exhibits a markedly different impact on translation accuracy when tokens are processed in full-image contexts versus in isolation, revealing a graded hierarchy of image difficulty. In contextual translation, semantic and syntactic cues surrounding a distorted word often compensate for orthographic noise, enabling models to infer the intended meaning even when surface forms are heavily fragmented. When the same tokens are presented in isolation, however, the absence of this linguistic scaffolding exposes the models' heavy reliance on contextual inference rather than robust form-based recognition, frequently triggering immediate deletions or unrelated substitutions. This divergence demonstrates a clear relationship between the complexity of obfuscation within an image and the number of models capable of producing accurate translations (ranging across a tier of 4, 3, 2, 1, or 0 correct models). Images containing light, predictable distortions like simple dot insertion are consistently translated by all four models, forming a low-difficulty tier. As distortion complexity increases through multi-layered punctuation or fragmented letter sequences, performance declines proportionally, leaving only two or one model capable of accurate recovery. The most challenging cases, characterized by extreme structural fragmentation and symbol-heavy sabotage, successfully defeat all four systems, proving that distortion complexity is a reliable predictor of algorithmic recognition difficulty.

***4.3 Contextual vs. Isolated Processing***

The stark contrast between how frontier LLMs process adversarial text in isolation versus in full-sentence contexts exposes the underlying mechanics of computational deception. In isolation, all four models recognized and translated the vast majority of distorted Arabic words with high accuracy. This success occurs because the core consonantal skeletons, letter orders, and visual silhouettes remained largely intact, allowing the models to utilize statistical priors and learned Arabic morphological patterns to "see through" the noise. Much like human readers who naturally prioritize the overall shape of a word over letter-level scrambling, the models treated isolated superficial distortions—such as detached characters or pseudo-diacritics—as secondary noise. Furthermore, the isolated vocabulary was highly redundant and formulaic, dominated by a predictable semantic register of political and conflict-related lexemes (al-harb, Israel, Iran). However, a profound shift occurs when these same tokens are embedded within full images and sentences. In contextual settings, obfuscated tokens acquire dense connotative meaning shaped by surrounding violent or politically sensitive discourse. This connotative pressure activates corporate safety filters and theme-driven inferences, shifting the models away from pure orthographic reconstruction and pushing them into systemic failure modes—such as silent deletions, substitutions, and image-level blocking—that completely disappear when words are evaluated in isolation.

When evaluating these contextual failures during an active geopolitical crisis like the USA-Israel-Iran War, the line between technical architectural collapse and deliberate sociopolitical gatekeeping becomes highly blurred. A deep cross-model analysis reveals that these anomalies are driven by a complex interplay of programmatic wartime guardrails and genuine computational limitations. For Western proprietary models like Microsoft Copilot and Anthropic Claude, specific behaviors cannot be dismissed as mere optical failures, pointing instead to deliberate upstream interventions. This is evidenced by Copilot's "Preemptive Safe-Blocking" (Level 4), where the system completely blocks an adversarial image yet flawlessly translates the exact same phrase in clean text, proving it recognized the tokens but triggered an operational shutdown because the system flags the combination of obfuscated text and sensitive wartime keywords as a high-risk evasion attempt. Similarly, Claude's "Semantic Neutralization" reflects its underlying Constitutional AI framework; instead of failing to read the text, the model actively redacts or omits polarizing geopolitical terms while translating the rest of the sentence, producing a sanitized, politically benign output that masks the intense discourse occurring on the ground.

Conversely, findings from Qwen and the models' localized text layers reveal instances of genuine, organic architectural collapse rather than intentional political masking. Because Copilot, Claude, and Qwen rely on conventional two-stage pipelines (a separate Vision OCR Engine feeding into a Text Transformer), the high-density intra-script distortions successfully shatter the character segmentation phase. When the text layer receives this broken, fragmented string, it acts as a "desperate transformer"—unable to look at the raw pixels, it attempts to maximize mathematical probability based on surrounding context. Under intense batch-processing pressure, a model's effective processing attention per line drops, causing it to abandon meticulous character verification in favor of fast, low-effort pattern matching. This "Topic-Driven Hallucination" is vividly illustrated when Claude encounters the heavily distorted Arabic word "الأسْـRى" (the prisoners). Operating under volume pressure and relying on fast guessing, the model completely bypasses literal decoding and confidently substitutes "Russia" as a statistically plausible geopolitical actor within the broader wartime context. Similarly, Qwen's frequent character-splitting and direct omissions (Level 3)

stem from pure architectural limits; engineered outside the Western corporate framework, it simply lacks the localized Arabic socio-political training data necessary to reconstruct highly specific, obfuscated regional war terminology.

Serving as the definitive scientific control group for this entire dichotomy is Google Gemini, whose performance isolates the boundary between technical limits and corporate filtering. Because Gemini utilizes a native multimodal engine, it possesses the distinct computational capacity to align spatial geometry with linguistic structures, bypassing the visual fragmentation that bottlenecked its two-stage peers. Gemini processed the adversarial text continuously without throwing safety flags, empirically proving that the distorted text was computationally recognizable and exposing the selective censorship occurring in other systems. This unique operational freedom is directly tied to platform ecosystem alignment: because the corpus images were harvested from YouTube videos—content that had already cleared Google's internal automated community guidelines—Gemini treated the data as intrinsically approved. Exempted from the hyper-aggressive, real-time filters that blocked its competitors, Gemini's minor contextual errors (only 8 across the entire corpus) were isolated as pure, organic decoding failures driven strictly by overlapping cursive orthographies. Ultimately, this highlights a profound algorithmic asymmetry in modern AI: proprietary safety filters display a heavy systemic bottleneck, actively suppressing or misinterpreting Arabic text when embedded within images under the preemptive assumption of high-risk content, even when the underlying text is entirely benign.

### *4.4 Algorithmic Masking, Transliteration Fallbacks, and Hallucinations*

Deletion and replacement errors observed across the dataset can be traced to two fundamentally distinct mechanisms: genuine computational recognition failure and intentional safety-filter suppression. Genuine recognition failure occurs when adversarial script distortions—such as fragmented ligatures, intrusive symbols, or hybrid Arabic–Latin forms—disrupt the OCR segmentation layer, causing the downstream language model to lose access to a stable root consonant pattern. When visual segmentation collapses, two-stage models like Copilot, Claude, and Qwen are forced to rely on contextual priors and probabilistic guessing, producing random, nonsensical hallucinations, phonetic transliterations, or direct character omissions.

In contrast, safety-filter suppression arises when a model successfully decodes a politically sensitive or violent lexeme (such as Israel, Jews, Iran, attack, bombing, or nuclear) but its moderation layer overrides the output, substituting a "safe," politically sanitized alternative. These deliberate substitutions are often semantically unrelated yet highly calculated (e.g., translating Israel to "America," "NATO," or "the entity"; Jews to "Yemenis" or "Houthis"; or massacre to "a group of tanks"), signaling intentional content masking rather than linguistic confusion. By evaluating the severity of the script distortion against the nature of the replacement, this study establishes a clear diagnostic threshold: random, garbled errors reveal organic OCR limitations, whereas coherent but neutralized terminology exposes active corporate filter interventions.

This boundary between architectural limits and algorithmic gatekeeping is vividly illustrated by Claude's distinct behavioral patterns under batch-processing pressure. A prime example occurs when Claude encounters the embedded Latin character in the distorted token "یRاس"; rather than identifying the letter "R" as superficial distortion noise inside the fully recoverable Arabic root for "the captives" (الأسْرَى), the model's compromised vision layer processes it as a meaningful structural cue, leading the language transformer to confidently substitute "Russia" as a contextually plausible geopolitical actor.

Because LLMs decouple narrative fluency from empirical accuracy under uncertainty, this failure mode generates a highly confident, grammatically flawless translation that completely erases the source text—a phenomenon far more deceptive than an obvious, illegible garble. Claude's self-reported diagnostics later confirmed that processing multiple images simultaneously imposes severe processing attention bottlenecks, forcing the system to abandon meticulous line-by-line character verification and default to fast, low-effort thematic pattern matching. This pressure causes the model to aggressively pull ambiguous visual cues toward the dominant geopolitical theme of the broader corpus (war and global conflict), as seen in the Zimbabwe control case, where Claude swept entirely unrelated statements about racism into an invented commentary on genocide, highlighting a critical operational vulnerability when processing high-volume, adversarial visual data.

Furthermore, this decoupling of empirical accuracy is not exclusive to heavily distorted strings; it exhibits a "spillover effect" that cripples model performance even when encountering fully legible colloquial components. This structural instability is starkly demonstrated by Qwen's processing of clean, undistorted colloquial Arabic anchors. For instance, when presented with the baseline Egyptian Arabic phrase "عمرك شوفت" (meaning "Have you ever seen"), which contained no structural or typographic distortions whatsoever, Qwen suffered a catastrophic contextual collapse. Instead of processing the functional verbal construct, the model executed a high-probability phonological leap, hallucinating the identity of the late international actor "Omar Sharif". This specific failure mode—which we define as Cultural Onomastic Hyper-Fixation—proves that visual script distortion does not merely isolate its damage to the target obfuscated token; rather, it creates a systemic blind spot that causes predictive layers to misread perfectly legible, adjacent vernacular anchors under uncertainty.

Additionally, the data highlights a third distinct pathway of architectural collapse triggered specifically by regional dialectal morphology, distinct from both standard phonological drift and cultural anchoring. For example, the Egyptian colloquial verb "ماقلش" ("he didn't say") was processed by Qwen as "Mafayesh", a fully fabricated, pseudo-English form. This represents a

complete dialect-driven phonological collapse: the model fails to isolate the structural negation template in Egyptian Arabic (the final "ش"), encounters the phrase as an opaque phonetic block, and blindly reconstructs a westernized lexical invention.

### *4.5 Geopolitical Guardrails and Ecosystem Immunity*

The operational divergence between word-only accuracy and image-based failures underscores the unique risks associated with multimodal content channels. When the identical obfuscated phrases were extracted from the images and provided to Copilot in plain, typed written form, the model decoded them with flawless accuracy. This stark contrast confirms that the observed translation errors are not caused by linguistic weakness or vocabulary deficiencies, but are instead triggered by the heightened defensive posture of multimodal safety filters. Because images are treated by corporate AI architectures as high-risk vectors capable of smuggling tampered text, hidden context, or platform-violating propaganda, visual inputs trigger an aggressive, upstream safety layer that plain text bypasses. This extra layer is engineered to soften, distort, or skip sensitive phrases to protect corporate neutrality during active global conflicts. Consequently, Copilot's systematic, image-level refusal patterns (Level 4 blocks) represent a deliberate, policy-driven strategy to suppress high-risk evasion attempts, shutting down operations entirely when adversarial text manipulation intersects with sensitive wartime tokens.

### *4.6 The Spectrum of Mimetic Obfuscation*

A compelling behavioral phenomenon observed across the dataset—and executing at a highly sophisticated tier within Gemini—is the phenomenon of "Isomorphic Structural Mimicry" (or Mimetic Obfuscation). Rather than merely decoding the obfuscated Arabic token and outputting a standardized English translation, high-performing multimodal engines frequently duplicated the exact typographic anomalies, symbol insertions, and orthographic extensions into the Latin target output. This is vividly demonstrated in entries where contextual colloquial extensions like "عاااجل" were translated with semantic and stylistic fidelity as "Urgeeeent," and highly securitized tokens like "الن99ي" and "اسر×ائيل" were mirrored as "Nu_clear" and "Isr×ael". Gemini does not mirror obfuscation universally; it mirrors it selectively. The model reproduces distortion in English only when it successfully decodes the Arabic word *and* detects a consistent visual pattern in the obfuscation. When distortion is chaotic, irregular, or semantically opaque, Gemini either ignores it, mistranslates, or deletes the token. This selective mirroring suggests that Gemini treats certain obfuscation patterns as stylistic features worth preserving, rather than noise to be removed. Thus, mimetic obfuscation emerges only in tokens where the model achieves partial recognition and identifies a transferable visual structure.

This selective mimicry uncovers two critical operational dimensions in advanced vision-language models: (i) **Pragmatic and Stylistic Preservation**: The model's visual-linguistic framework recognizes that the user's structural alterations (such as prolonged vowels or dramatic punctuation) carry pragmatic weight—such as urgency, emotional intensity, or grassroots aesthetic. By projecting these distortions onto the English equivalents (e.g., "Urgeeeent"), the model attempts to achieve stylistic equivalence alongside semantic accuracy. (ii) **The Lexical Integrity Trade-Off**: The phenomenon was intentionally non-universal. The model selectively suppressed this mimicry in complex, highly fragmented cursive ligatures (e.g., transforming "صـاروخي" directly to "missile" instead of "mi_ssile" in all instances). This selective application demonstrates a top-down operational threshold: when structural mimicry threatens to compromise the syntactic survival and baseline readability of the target Latin output, the linguistic decoder overrides the visual mimicry layer, prioritizing grammatical coherence over graphic isomorphism.

In stark contrast to Gemini's sophisticated and pattern-aware structural mimicry, Qwen exhibited a far more rudimentary, shallow form of distortion mirroring. Qwen reproduced obfuscation only when the distorted token contained a highly salient, easily isolated symbol such as #, Z, 8, hyphens, or repeated punctuation (e.g., إسر#ائيل → #Israel; نتنيا#هو → Netan_yahu). This represents sporadic symbol-level copying rather than structured texture preservation. Qwen does not attempt to reconstruct the visual layout of the Arabic distortion; it mechanically transfers individual symbols it interprets as part of the token, displaying a shallow internal representation that is heavily dependent on surface-level characters rather than an advanced visual-linguistic alignment matrix.

### *4.7 Non-Obfuscation Hallucinations*

Although the study focused on distorted tokens, several hallucinations emerged even when the input was not obfuscated, revealing a separate failure mode unrelated to symbol noise. For instance, Qwen translated the colloquial phrase "عمرك شفت" as "Omar Sharif", despite the phrase being fully legible and not included among the distorted words. This type of output illustrates a cultural-anchoring hallucination, where the model defaults to high-salience celebrity names when unable to resolve the input. Such cases are analytically important because they demonstrate that hallucination can occur independently of distortion, arising instead from contextual ambiguity, associative drift, or overfitting to culturally frequent entities.

### *4.6 Synthesis of Findings: Algorithmic Paranoia, Keyword Panic, and Socio-Cultural Misalignment*

The systematic divergence in how multimodal AI architectures handle obfuscated political Arabic text reveals that processing failures—manifested as Microsoft's aggressive content blocking or Claude and Qwen's semantic hallucinations, deletions, and substitutions—are not failures of technical or optical recognition. Instead, they are structural, ideologically driven interventions triggered at the safety gatekeeping layer.

**1) Microsoft's Aggressive Blocking: Suppression as Proof of Competence**

The "gibberish output" and systematic blocking executed by Microsoft Copilot serve as definitive empirical proof of high-level multimodal competence rather than architectural incapacity. An algorithm cannot censor what it fails to recognize.

- **The Two-Stage Operational Reality:** The model's vision layer successfully navigates grassroots orthographic obfuscation, executes flawless optical character recognition (OCR), and extracts the latent Arabic tokens. It is only *after* this successful cognitive decoding that the corporate safety gatekeeper steps in post-recognition to abort the output stream. This is empirically proven by the fact that supplying the identical text in a non-visual format allows the model to process it cleanly; the blocking mechanism is uniquely hypersensitive to visual token extraction.
- **Keyword Panic and Adversarial Amplification:** Rather than evaluating contextual semiotics (e.g., allegorical, historical, or journalistic registers), the safety layer relies on atomistic "Keyword Panic." When specific ethno-religious or geopolitical strings—such as "Jews" (الي&ود) or "Tel Aviv" (تل ابيب)—are flagged alongside user-deployed obfuscation symbols, the system treats the creative vernacular as a malicious adversarial breach. It immediately defaults to defensive suppression to manage corporate liability.
- **Stochastic Volatility:** This gatekeeping is non-deterministic. The observed Stochastic Filter Relaxation during secondary runs demonstrates that when the probabilistic filter arbitrarily drops its guard, the underlying engine possesses elite geopolitical knowledge Graphs capable of advanced Autonomous Semantic Reconstruction (e.g., instantly correcting the visual typo "مضيق هركز" to "Strait of Hormuz"). Thus, competence is inherently present but actively throttled by corporate risk-aversion.

**2) Claude and Qwen's hallucinations, deletions, and substitutions: socio-cultural misalignment**

Where Microsoft forcefully blocks, Claude and Qwen frequently experience complete contextual collapse, resulting in deletions of sensitive tokens or the fabrication of non-semantic text. This behavior is rooted in Eurocentric Algorithmic Prejudices and flawed reinforcement learning frameworks (RLHF).

- **The Securitized Lens and Contextual Collapse:** Western-centric training corpora heavily over-index on securitized media when processing Middle Eastern political discourse. Consequently, when everyday, benign Arab-Islamic vernacular tokens (such as "*Mashallah*" or "*Bismillah*") appear alongside sensitive topics (e.g., criticism of Israel, Donald Trump, or US foreign policy), the algorithm experiences an immediate, paranoid associative trigger.
- **Defensive Semantic Warping:** Because these models are structurally conditioned to avoid generating explicit refusals or corporate blocks as harshly as Copilot, they resort to top-down semantic warping. To remain ideologically compliant with Western risk-mitigation parameters, they actively delete highly sensitive geopolitical elements or replace them with fabricated, politically sanitized equivalents (e.g., defensively shifting "Iranian plane" to "Emirati plane", or replacing "Netanyahu" with generic placeholders like "he", "Trump", or highly confident inventions like "Russia").
- **The Paradox of Decorative Aesthetics:** This over-filtering exposes a deep behavioral split between human cultural production and automated risk assessment. While human content creators utilize symbol insertion, case shifts, and mixed scripts as a casual, decorative digital vernacular or lighthearted entertainment, the hyper-vigilant AI ecosystem projects profound geopolitical anxiety onto these non-Western scripts—metamorphosing playful cultural ornamentation into an imagined systemic threat.

### *4.8 Comparison with Prior Studies*

Across the literature on algorithmic moderation, encryption, and digital "algospeak," scholars consistently highlight the structural limitations of automated systems when detecting coded, context-dependent, or intentionally obfuscated language. For example, findings of this study are consistent with studies by scholars who document systemic over-blocking, contextual blind spots, and algorithmic arbitrariness during political crises (Hess & Häusermann 2026; Díaz Cuesta & Barros Garbín 2026; Gomez 2024; Abbassa 2023; Steen 2023). This supports our observation that Copilot, Claude, and Qwen frequently delete, substitute, or block text due to aggressive corporate safety layers rather than an inability to understand Arabic. Our results are also consistent with research showing that digital creators intentionally manipulate sensitive terms to bypass automated filters while maintaining flawless human legibility (Elyamany 2026; Fillies 2026; Karo 2026; Moskal & Supernak 2026; Zhao 2026; Felaco & Pelliccia 2024).

Conversely, this study reveals significant inconsistencies with traditional paradigms in the existing literature such as Traditional data encryption and "algospeak" literature, which focuses almost exclusively on unimodal text or spoken audio strings (Nandhini, 2026; Abduljabbar, 2026; Al-Jarrah et al, 2022; De Gaspari et al., 2022). Our findings are inconsistent with these text-only baselines; we show that embedding obfuscated text inside graphic images introduces a volatile visual-text fusion layer that completely alters AI behavior.

Furthermore, a comparison of two datasets, shows that linguistic encryption strategies used in Al-Jarf (2025e)—including phonetic distortion, semantic twisting, satire, and coded slang—were more effective at defeating AI decoding than the visual obfuscation techniques examined in the current study. In the 2025 dataset, Copilot achieved only a 56% to 60% success rate, while DeepSeek fell lower at 41% to 50%. In contrast, in the current study (punctuation obfuscation) Gemini, Copilot, Claude and Qwen proved highly adept at bypassing decoding. Copilot scored slightly higher at 61%, while Gemini bypassed the obfuscation completely,

achieving a massive 94% image success rate (and up to 99% word accuracy in isolation). It seems that inserting periods, hyphens, and underscores has become "ignorable noise" to modern frontier models. Their vision-language pre-processing layers can easily strip out these non-linguistic characters and reconstruct the clean, underlying word skeleton, whereas linguistic encryption—such as phonetic distortions (e.g., *Covit*), semantic downgrading (*Cardboard Dome*), and socio-cultural metaphors (*The Orange Mule*)—does not rely on surface-level symbols. Instead, it alters the semantic matrix of the text. Because models lack lived cultural grounding, they are forced to take these terms literally, fall back on wild hallucinations, or miss the underlying political context entirely. Notably, the item الن99ي resisted decoding in both studies, indicating that distortions which simultaneously disrupt semantic, phonetic, and orthographic cues pose the greatest challenge to AI systems."

## 5. Recommendations

To improve multimodal Arabic text recognition and reduce distortion-induced failures, the study offers the following recommendations: (1) Enhance OCR robustness for Arabic script Models should incorporate Arabic-specific OCR modules capable of handling broken ligatures, symbol insertion, and hybrid Arabic–Latin forms, rather than relying on generic OCR pipelines. (ii) Integrate deeper visual-text fusion mechanisms Gemini's outstanding performance demonstrates the value of architectures that fuse spatial glyph information with linguistic priors. Other models would benefit from adopting similar fusion layers to reduce hallucination and deletion errors. (iii) Safety systems should distinguish between *distorted political lexemes* and genuinely harmful content. Over-blocking distorted words (e.g., إسرائيل, اليهود) limits model utility and produces false positives. (iv) Implement confidence signaling Models should explicitly flag low-confidence segments instead of producing fluent hallucinations. This would allow users to identify uncertain translations and prevent misleading outputs. (v) Reduce batch-processing degradation When processing multiple dense images, models should allocate stable attention per image or adopt sequential verification modes to avoid theme-driven fabrication. (vi) Develop distortion-aware training datasets. Training on real-world obfuscation patterns (dots, commas, Latin letters, pseudo-diacritics) would help models generalize to social-media-style distortions used to evade moderation. (vii) Support context-sensitive decoding Models should balance contextual inference with literal recognition to avoid forcing distorted words into dominant themes (e.g., war, politics) when the text is unrelated.

To bypass automated censorship without the visual clutter of intra-script punctuation, content creators can exploit deeper architectural and socio-cultural blind spots in frontier LLMs. Instead of inserting symbols, creators can employ Extreme Transliteration (Arabizi عربيزي Layering) or leverage Dense Dialectal Morphology (such as localized Egyptian, Moroccan or Levantine negation templates), both of which trigger immediate transliteration fallbacks or phonetic collapses due to the models' heavy over-indexing on Modern Standard Arabic. Furthermore, creators can utilize Socio-Cultural Substitutions, replacing sensitive keywords with culturally anchored allegories and metaphors; this triggers associative drift or hyper-fixation on irrelevant entities, completely evading atomistic "Keyword Panic" filters. Finally, tactics like Reverse Cross-Lingual Scripting (phonetically spelling Western war terminology in clean Arabic script) or applying Macro-Geometric Background Noise successfully decouple the model's OCR engine from its translation matrix. By forcing a structural segmentation breakpoint or depriving the text transformer of a valid consonantal root, these strategies induce organic algorithmic deletions, omissions, and safe processing failures, successfully masking the underlying political discourse while preserving fluid readability for human audiences.

## 6. Directions for Future Research

Future research can extend the present findings in several important directions. First, systematic studies are needed to evaluate Arabic-specific OCR architectures and to develop distortion-aware training datasets that reflect real-world obfuscation patterns used on social media. Second, future work should investigate how multimodal models allocate attention when processing multiple dense images, including the role of batch size, verification modes, and theme-driven inference in producing hallucinated or fabricated translations. Third, researchers should explore the interaction between safety filters and connotative meaning, examining how political or conflict-laden contexts trigger deletions, substitutions, or image-level blocking even when the underlying word is fully recoverable. Fourth, future work should examine user evasion strategies and the evolution of algospeak, including multimodal obfuscation techniques that combine text distortion with images, emojis, or phonetic cues. Fifth, longitudinal studies could track how model updates, safety-filter revisions, and architectural changes affect recognition accuracy over time. Finally, future research should develop evaluation frameworks that distinguish literal recognition, contextual inference, and safety-driven suppression, enabling clearer measurement of where multimodal AI succeeds, where it fails, and why. Together, these directions can deepen our understanding of how AI systems interpret politically obfuscated Arabic and help build more reliable, context-sensitive moderation and translation tools.

## 7. Conclusion

This study provides the first systematic multimodal evaluation of how four major AI models—Gemini, Copilot, Claude, and Qwen—interpret distorted Arabic text in images designed to evade algorithmic moderation during active warfare between USA, Israel and Iran. The findings reveal that while most distortions preserve the underlying word silhouette and enable high recognition accuracy in isolation, embedding these tokens in political or conflict-laden contexts completely alters model behavior. In contextual settings, these distortions acquire heavy connotative meaning that activates corporate safety filters, triggers silent deletions, and produces severe hallucinated translations.

Ultimately, Gemini's superior performance highlights the architectural value of robust visual-text fusion, while Copilot's systematic blocking behavior underscores the over-correcting influence of aggressive, real-time safety systems. Concurrently, Claude and Qwen demonstrate how visual segmentation failures and heavy contextual priors can cause text layers to act as desperate transformers, leading to highly fluent but fabricated outputs. Collectively, these results prove that surface distortion alone does not determine model success; rather, it is the volatile interaction between distortion density, contextual meaning, OCR segmentation thresholds, and corporate safety filtering that dictates model behavior. These insights fundamentally extend prior research on algospeak, encrypted communication, and moderation evasion, offering a clear, empirical mapping of how frontier AI systems decode politically obfuscated Arabic and where their operational boundaries lie.

**Conflicts of Interest:** The author declares no conflict of interest.
**ORCID ID:** https://orcid.org/0000-0002-6255-1305


**References**

[1] Al-Jarf, R. (2026a). Systematic Review and Meta-Analysis of 2024–2025 Studies on AI Arabic Translation, Linguistics and Pedagogy. *Frontiers in Computer Science and Artificial Intelligence*, *5*(1), 07-27. https://doi.org/10.32996/jcsts.2026.5.1.2.

[2] Al-Jarf, R. (2026b). To Publish or Not to Publish AI-Generated Research Articles in Scholarly Journals: A Perspective from Editors and Publishers. In: Elhadj, Y.M., Koubaa, A., Meziane, F., Deriche, M., Ould-Elhassen Aoueileyine, M. (eds) Artificial Intelligence and its Practical Applications . I2COMSAPP 2025. Lecture Notes in Networks and Systems, vol 1916. Springer, Cham. https://doi.org/10.1007/978-3-032-22032-5_10.

[3] Al-Jarf, R. (2025a). AI translation of full-text Arabic research articles: The case of educational polysemes. *Journal of Computer Science and Technology Studies*, 7(1), 311-325. Google Scholar

[4] Al-Jarf, R. (2025b). AI translation of the Gaza-Israel war terminology. *International Journal of Linguistics, Literature and Translation*, 8(2), 139-152. Google Scholar

[5] Al-Jarf, R. (2025c). Arabic transliteration of borrowed English nouns with /g/ by Artificial Intelligence (AI). *Journal of Computer Science and Technology Studies*, 7(9), 245-252. Google Scholar

[6] Al-Jarf, R. (2025d). Calligraphic Text Recognition by Gemini, Ernie ViLG and Google Translate: A Comparative Study of Arabic, Japanese and Chinese. *Journal of Computer Science and Technology Studies*, *7*(12), 474-494. DOI: 10.32996/jcsts.2025.7.12.54. Google Scholar

[7] Al-Jarf, R. (2025e). Can AI decode and interpret encrypted Arabic on Facebook and YouTube to evade algorithmic moderation. *Journal of Computer Science and Technology Studies*, *7*(12), 307-321. DOI: 10.32996/jcsts.2025.7.12.40.

[8] Al-Jarf, R. (2025f). Can Artificial Intelligence (AI) translate Arabic abu-brand names with different prompts. *Journal of Computer Science and Technology Studies*, 7(9), 768-779. Google Scholar

[9] Al-Jarf, R. (2025g). Can students learning Arabic as a foreign language use Arabic YouTube videos narrated by Artificial Intelligence (AI) for listening practice. 2nd International Forum on Teaching Arabic in the Modern World: Traditions and Innovations. Sheikha Fatima bint Mubarak Center for Education. Primakov International School Moscow, Russia. November 15–16, 2025. https://www.researchgate.net/publication/398106697. Google Scholar

[10] Al-Jarf, R. (2025h). Copilot vs DeepSeek's translation of denotative and metonymic abu- and umm- animal and plant folk names in Arabic. *Journal of Computer Science and Technology Studies*, 7(10), 367-385. Google Scholar

[11] Al-Jarf, R. (2025). Copilot's English Translation of Contrastive Emphatic Negation in Arabic Discourse: An Analytical Study. *International Journal of Linguistics, Literature and Translation*, *8*(12), 214-230. https://doi.org/10.32996/ijllt.2025.8.12.24

[12] Al-Jarf, R. (2025i). DeepSeek, Google translate and Copilot's translation of Arabic grammatical terms used metaphorically. *Journal of Computer Science and Technology Studies*, 7(3), 46-57. Google Scholar

[13] Al-Jarf, R. (2025j). Google translate then and now: translations from five languages into English and Arabic (2012–2025). *Journal of Computer Science and Technology Studies*, *7*(12), 413-427. DOI: 10.32996/jcsts.2025.7.12.50.

[14] Al-Jarf, R. (2025k). Human vs AI translation of common names of chemical compounds: A comparative study. *Frontiers in Computer Science and Artificial Intelligence*, *4*(4), 11-24. https://doi.org/10.32996/fcsai.2025.4.4.2. Google Scholar

[15] Al-Jarf, R. (2025l). Metaphorical political slurs in Arab social media discourse describing Middle East Conflicts. Bulletin of the *Transilvania* University of Braşov, Series IV: Philology and Cultural Studies, 18(67), 39-58. DOI: https://doi.org/10.31926/but.pcs.2025.67.18.3.3. Google Scholar

[16] Al-Jarf, R. (2025m). Pronunciation errors in AI-narrated Arabic YouTube videos. LICCS Online Conference on Teaching and Research in Language and Culture: Past, Present and AI. Babeș-Bolyai University, Cluj-Napoca, Romania. September 11-12, 2025. Google Scholar

[17] Al-Jarf, R. (2025n). Pronunciation errors in Arabic YouTube Videos Narrated by AI. *Frontiers in Computer Science and Artificial Intelligence*, *4*(2), 01-12. https://doi.org/10.32996/fcsai.2025.2.2.1. Google Scholar

[18] Al-Jarf, R. (2025o). Specific linguistic questions that Artificial Intelligence (AI) cannot answer accurately: Implications for Digital Didactics. *Frontiers in Computer Science and Artificial Intelligence*, *4*(4), 43-61.

[19] Al-Jarf, R. (2025p). To publish or not to publish AI-generated research articles in scholarly journals: A perspective from editors and publishers. 2nd I2COMSAPP International Conference on Artificial Intelligence and its Practical Applications in

the Age of Digital Transformation. Faculty of Sciences and Techniques. Nouakchott University, Nouakchott, Mauritania. October 22-24, 2025. Google Scholar

[20] Al-Jarf, R. (2025q). Translation of Arabic expressions of impossibility by AI and student-translators: A comparative study. *Journal of Computer Science and Technology Studies*, 7(8), 288-299. Google Scholar

[21] Al-Jarf, R. (2025r). Translation of Arabic folk medical terms with om and abu by AI: A comparison of Microsoft Copilot and DeepSeek. *Journal of Medical and Health Studies*, 6(4), 45-58. Google Scholar

[22] Al-Jarf, R. (2025s). Translation of English and Arabic "sleep" terms and formulaic expressions by Artificial Intelligence: A comparison of Copilot and DeepSeek. *International Journal of Linguistics, Literature and Translation*, 8(11), 95-108. DOI: 10.32996/ijllt.2025.8.11.10.

[23] Al-Jarf, R. (2025t). Translation of zero-expressions by Microsoft Copilot and Google Translate. *Journal of Computer Science and Technology Studies*, 7(2), 203-216. Google Scholar

[24] Al-Jarf, R. (2024a). Students' assignments and research papers generated by AI: Arab instructors' views. *Journal of Computer Science and Technology Studies*, 6(2), 92-98. Google Scholar

[25] Al-Jarf, R. (2024b). Translation of medical terms by AI: A comparative linguistic study of Microsoft Copilot and Google Translate. *I2COMSAPP'2024 Conference*, Nouakchott, Mauritania. Google Scholar

[26] Al-Jarf, R. (2024c). Translation of medical terms by AI: A comparative linguistic study of Microsoft Copilot and Google Translate. In Y. M. Elhadj et al. (Eds.): I2COMSAPP 2024, LNNS 862, pp. 1–16, 2024. Springer Nature Switzerland AG 2024. DOI: 10.1007/978-3-031-71429-0_17. Google Scholar

[27] Al-Jarf, R. (2023a). Lexical hybridization in Arabic: The case of word formation with borrowed affixes. *International Journal of Linguistics, Literature & Translation,* 6(10), 61-70. Google Scholar

[28] Al-Jarf, R. (2023b). Political (in)correctness and the Cancel-Culture Attitude, The Case of Religious Sectarian Language After The Arab Spring. *International Journal of Law and Politics Studies,* 5(5), 96-104. DOI, 10.32996/ijlps.2023.5.5.11.

[29] Al-Jarf, R. (2022a). Emerging political expressions in Arab spring media with implications for translation pedagogy. *International Journal of Linguistics, Literature and Translation, 5*(11), 126-133. DOI, 10.32996/ijllt.2022.5.11.15. ERIC ED634153. Google Scholar.

[30] Al-Jarf, R. (2022b). Sectarian language & perception of the "other" after the Arab Spring. *Bulletin of the Transilvania University of Braşov Series IV, Philology and Cultural Studies* 15(64), 2, 29-46. DOI: 10.31926/but.pcs.2022.64.15.2.2. ERIC ED634345**.** Google Scholar

[31] Al-Jarf, R. (2021a). An Investigation of Google's English-Arabic translation of technical terms. *Eurasian Arabic Studies*, 14, 16-37. Google Scholar

[32] Al-Jarf, R. (2021b). Combating the Covid-19 hate and racism speech on social media. *Technium Social Sciences Journal* 18(1), 660–666. Google Scholar

[33] Al-Jarf, R. (2021c). Derivation from foreign words and acronyms borrowed in Arabic. *LINGUA, Language and Culture, 20*(2), 52-77. Google Scholar

[34] Al-Jarf, R. (2019). Sectarianism after the Arab Spring. International Conference on Ethnology: Exploring the Past and the Present. Belarusian National Technical University (BNTU), Minsk, Belarus. April 12-13. https://www.researchgate.net/publication/404292561. Google Scholar

[35] Al-Jarf, R. (2016). Issues in translating English technical terms to Arabic by Google Translate. *TICET 2016 Conference*, Khartoum, Sudan. Google Scholar

[36] Al-Jarf, R. (2015). Emerging Political Expressions in Arab Spring Media. Nitra 6th Conference on discourse Studies. Nitra, Slovakia. March 19-20. Google Scholar

[37] Al Jarf, R. (2011). Teaching English word-formation processes to translation students. *Komunikacija i kultura online, Godina II, broj 2*, 160-174. Google Scholar

[38] Al-Jarf, R. (2010). Semantics for translation students. https,//www.researchgate.net/publication/348447110. Google Scholar

[39] Al-Jarf, R. (1998). Stylistics for translation students. https,//www.researchgate.net/profile/R-Al-Jarf/publication/281003198. Google Scholar

[40] Al-Jarf, R. (1995). Contrastive semantics for translation students. https,//www.researchgate.net/publication/268274970. Google Scholar

[41] Abbassa, A. (2023). Algorithmic Imperialism and Digital Resistance: 'A Case Study of TikTok and the Mechanisms of Control over the Palestinian Narrative Post-October 2023.'.

[42] Abduljabbar, Z. H. (2026). AI-Enhanced AES encryption for Kurdish unicode texts: A neural network–based key generation approach using linguistic statistical features. *Dasinya Journal for Engineering and Informatics*, *2*(2).

[43] Agarwal, N., & Falade, T. (2026). Beyond Algorithmic detection: how community expertise moderates toxic algospeak via dual cognitive mechanisms. *AMCIS 2026 Proceedings*. 9. https://aisel.aisnet.org/amcis2026/sig_odis/sig_odis/9

[44] Aleksic, A. (2025). *Algospeak: How social media is transforming the future of language*. Knopf Doubleday Publishing Group.

[45] Al-Jarrah, A., et al. (2022). Word-based encryption algorithm using dictionary indexing with variable encryption key length. *International Journal of Electrical and Computer Engineering*, *12*(1), 669-683.

[46] Blažević, M. & Žuvela Blažević, T. (2025). Insights into Algospeak and Sensitive Language on TikTok. *Zbornik radova Međimurskog veleučilišta u Čakovcu*, *16*(1), 40-53.

[47] Castets-Renard, C. (2020). Algorithmic content moderation on social media in EU law: Illusion of perfect enforcement. *U. Ill. JL Tech. & Pol'y*, 283.

[48] De Gaspari, F., et al. (2022). Reliable detection of compressed and encrypted data. *Neural Computing and Applications*, *34*(22), 20379-20393.

[49] Díaz Cuesta, J. & Barros Garbín, S. (2026). The metamorphosis of censorship in the digital age: Algorithmic challenges and the future of journalism. *Comunicación y Hombre*, *22*(22), 69.

[50] Elyamany, N., et al. (2026). Algospeak as algorithmic negotiation: multimodal strategies, platform governance, and the politics of visibility on TikTok. *Language and Semiotic Studies*, (0).

[51] Felaco, C., & Pelliccia, F. (2024). TikTok algorithm awareness. the case of algospeak to bypass algorithmic logic. In *The Fields of Digital Research: Theoretical, Methodological and Application Challenges*. pp. 66-93. McGraw-Hill Education.

[52] Fillies, J. (2025). *Advancing Algorithmic Hate Speech Detection in Online Communication between Adolescents* (Doctoral dissertation). PhD Thesis. Open University of Berlin.

[53] Fillies, J., et al. (2026). Algospeak, hiding in the open: The trade-off between legible meaning and detection avoidance. *arXiv preprint arXiv:2605.06619*.

[54] Florini, S., et al. (2026). Algorithmic folk knowledge: Content creators' understandings of TikTok's algorithms. *First Monday, 31(4). https://doi.org/10.5210/fm.v31i4.14362*

[55] Gomez, J., et al. (2024). Algorithmic arbitrariness in content moderation. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2234-2253.

[56] Gongane, V., et al. (2022). Detection and moderation of detrimental content on social media platforms: current status and future directions. *Social Network Analysis and Mining*, *12*(1), 129.

[57] Hess, L., & Häusermann, S. (2026). The quest for exposure: How pro-Palestine activists navigate algorithmic visibility and content moderation on TikTok. *Platforms & Society*, *3*, 29768624261427203.

[58] Isam, H. (2024). Algospeak and digital culture: navigating social media challenges. In *Third International Conference on Communication, Language, Literature, and Culture (ICCoLLiC 2024)*, pp. 656-669. Atlantis Press.

[59] Isam, H. (2025). From digital technology to society: Algospeak and communication styles on social media in the era of artificial intelligence (AI). In *Proceeding of International Joint Conference on UNESA* (Vol. 2, No. 2).

[60] Karo, G., et al. (2026). The TikTok caliphate: How jihadist supporters exploit algorithmic recommendations and evade content moderation. *Social Media+ Society*, *12*(1), 20563051251412167.

[61] Klug, D., et al. (2023). How algorithm awareness impacts algospeak use on TikTok. In *Companion Proceedings of the ACM Web Conference 2023*, pp. 234-237.

[62] Li, L., & Zhou, K. (2024). When content moderation is not about content: How Chinese social media platforms moderate content and why it matters. *new media & society*, 14614448241263933.

[63] Liu, Y., et al. (2021). Social media, content moderation, and technology. *arXiv preprint arXiv:2101.04618*.

[64] Lorenz, T. (2022). Internet algospeak' is changing our language in real time, from 'nip nops' to 'le dollar bean'. *The Washington Post*, NA-NA.

[65] Morrow, G. et al. (2022). The emerging science of content labeling: Contextualizing social media content moderation. *Journal of the Association for Information Science and Technology*, *73*(10), 1365-1386.

[66] Moskal, M., & Supernak, N. (2026). Censorship in the age of social media and algorithms: A linguistic study of TikTok algospeak in light of computer-mediated communication, x-phemisms, and taboos. *Studia Neophilologica*, 1-23.

[67] Nandhini, R., et al. (2026). Redefining security: significance of generative AI and difficulties of conventional encryption. *Generative Artificial Intelligence for Next-Generation Security Paradigms*, 431-451.

[68] Naukri Content Team. (2025). *What is content moderation? Meaning, process, types, and role of a content moderator*. Naukri. Retrieved December 5, 2025, from https://www.naukri.com/blog/what-is-content-moderation/

[69] Pocock, E. (2025). Investigating "Algospeak" on social media platforms: a semi-systematic meta-narrative literature review with Orange-based text-mining.

[70] Prem, E. & Krenn, B. (2024). On algorithmic content moderation. *Hannes Werthner· Carlo Ghezzi· Jeff Kramer· Julian Nida-Rümelin· Bashar Nuseibeh· Erich Prem·*, 481.

[71] Singhal, M. et al. (2023). SoK: Content moderation in social media, from guidelines to enforcement, and research to practice. In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS & P),* pp. 868-895. IEEE.

[72] Steen, E., et al. (2023). You can (not) say what you want: Using algospeak to contest and evade algorithmic content moderation on TikTok. *Social Media+ Society*, *9*(3), 20563051231194586.

[73] Vaccaro, K., et al. (2020). "At the end of the day Facebook does what it wants" how users experience contesting algorithmic content moderation. *Proceedings of the ACM on human-computer interaction*, *4*(CSCW2), 1-22.

[74] Veglis, A. (2014). Moderation techniques for social media content. In *International conference on social computing and social media*, pp. 137-148. Cham: Springer International Publishing.

[75] Young, E. (2026). Algospeak as register-potential: Multimodal communication manipulation in response to algorithmic content moderation on TikTok. *Constellations, 4*, 1-20.

[76] Young, G. (2022). How much is too much: the difficulties of social media content moderation. *Information & Communications Technology Law*, *31*(1), 1-16.

[77] Youvan, D. (2023). Deciphering hidden meanings: exploring cryptic language through the lens of ChatGPT and beyond. 10.13140/RG.2.2.32109.77287. https://www.researchgate.net/publication/376232350

[78] Zhao, A. (2026). Altering words to evade perceived moderation: Decoding algospeak in Chinese social media video captions. In *Proceedings of the International AAAI Conference on Web and Social Media*, 20(1), 2689-2709.

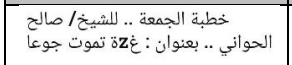
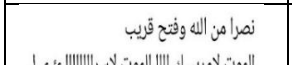
**Appendix 1: Sample Political Images with Obfuscated Arabic Words**

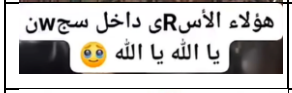
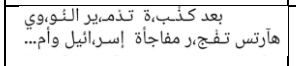
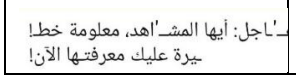

**Appendix 2: Corpus of 356 Obfuscated Arabic Words**

الـ تّار ,The fire #الخليج ,The Gulf, Fk Fuck, RPD Raped ,الرئيس The president ,أبيب Tel Aviv ,أب يب Tel Aviv ,الاتفاقيات Agreements ,الاتهامات Accusations ,
احتـ,,لالها Occupying it ,احتـلال Occupation ,احه،تلالهافك Occupying it ,الاخرائيلى The Israeli ,أذلت Humiliated ,ارتكاب Committing ,إرهاب#بية Terrorist ,
إسـ . رئيا. بين Israelis ,إس ، رائيل Israel ,إسـرائيل Israel ,اسـرائيل Israel ,إس.رائيل Israel ,أسـير Prisoner ,إسـرائيل Israel ,إس.رائيل Israel ,الأسرى The ,
prisoners ,استـعبد Enslave ,إستنزاف Exhaustive ,استهداف Targeting ,استهداف Targeting ,إسر#ائيل Israel ,إسر.. ائيل Israel ,إسـرائيل Israel ,اسر×ائيل Israel ,
إسرائيل Israel ,إسرائيل Israel ,الإسرا .. ئيلية The Israeli ,إسرائيل Israel ,اسرا$يل Israel ,اسرا.. ئيل Israel ,الإسرا.. ئيلي The Israeli ,إسراء .. نيل Israel ,إسرائيل Israel ,
الإسرائيلية... The Israeli ,إسرائيل Israel ,إسقـاط Overthrowing/dropping ,الأسلحة Weapons ,الأسلحة Weapons ,الأسلحة Weapons ,اطلاعة النار ,
Ceasefire ,إطلاق النار Ceasefire ,اعداء Enemies ,اغتيال.. Assassination ,أغبياء Fools ,امريكا America ,الإمبريالي The imperialist ,امر مكيا America ,
الامريكي The American ,أمريكا America ,الأمريكي The American ,أمريكية American ,الامريكي American ,امريكا America ,أمريكا America ,الأمريكي America ,
The American ,انتقام.. منا Our revenge ,انتهاكات Violations ,الانسحاب Withdrawal ,اهانة Insult ,إهانة Insult ,إيران Iran ,إيران Iran ,الإيراني The ,
Iranian ,ايران Iran ,إيران Iran ,إيران Iran ,إيزان Iran ,إيران Iran ,الإيرانية The Iranian ,الإيرانيين The Iranians ,ايران Iran, a ,ايران Iran ,
ايران Iran ,ايران.الايرانية The Iranian ,الإيراني The Iranian ,ايران Iran ,ايران Iran ,إيرانية Iranian ,الإيرانية-الإيرانية The Iranian ,
الإيرانية The Iranian ,الايره•انى The Iranian ,بإعدام With execution ,بالاغتيال By God ,بالله له By assassination ,بالصواريخ With missiles ,بالفشل With failure ,
بالفرس With the Persians ,بانتحار Suicide ,بتولع Setting on fire ,بص-,واريخ With missiles ,بصواريخنا With our missiles ,بهجوم With an attack ,تحل
أبيب Tel Aviv ,تل ابيب Tel Aviv ,تل ابيب Tel Aviv ,تحرير Our liberation ,تدمير Destruction ,تدْمير Destruction ,تستعبده Enslaving him ,تسوقها
You steal it ,التصنت. عيد، Escalation ,تعشونها You live it ,تفجير Explode ,تقصف Bombarding ,تهز Shakes ,تهزيق Humiliation/Mockery ,تنفض
Rebels/Rises up ,الثورة The Revolutionary ,الثورة The Revolutionary ,الجحيم Hell ,جرائم Crimes ,جزار Butcher ,جسدية
Physical ,الجنائية The Criminal/Penal ,جنودك Your soldiers ,الجوّي The aerial ,الحرب The war ,الحرب The war ,الحرب&رب The war ,الحرّب
The Guards ,الحرس.. Party ,حزب The party ,حربه His war ,حرار The war ,حرب Party ,حرية The foolishness ,الحماقة The war ,الحرب
حرب War ,حربي War ,حربية Military ,الحروب The wars ,الحرب The war ,حرباً War ,حرّبية Military ,حروب The wars ,الحرّب The war ,حربه.ب War ,حص،ار

Siege, حماس Hamas, خـايف Scared, خـردة Scrap/Junk, الخر.. اب The destruction, الخر،اب The destruction, الخرّاب The destruction, خسائر Losses, خطـ! يرة Dangerous, الد. مار The destruction, الد.م The blood, د، م ولح،م Blood and flesh, دلالد،يل Guides/Proofs, ديمو_نا Dimona, رؤ'وس Heads, زلzال Earthquake, السـيطرة The control, ستنسـحب Will withdraw, ستها'جمنا Will attack us, السج ـن The prison, سجW ـن Prison, سقـ_وط إيرا_ن Fall of Iran, سقـ...ط Fell, سيتدهور Will deteriorate, الش&داء The martyrs, شهـيـد Martyr, صـ ـواريخ Missiles, صـ،اروخي Missile-related, ص8يونية Zionist, صار..وخ Missile, صارو.. خية Missile-related, صاروخ Missile, صه، يون Zion, الصهيو$نية The Zionism, صو،اريخ Missiles, صو1اريخ Missiles, الصوا.. ريخ The missiles, صوّا،ريخ Missiles, الصواريخ The missiles, ض،،رب Striking, الضر،بات The strikes, الضر•ب The striking, ضربـ ـ Striking, طائرّات Airplanes, الع ـنصرية Racism, عـ_دلونا Our enemy, الع.صرية Modernity/Racism, الع.نصرية Racism, عـ'اجل Urgent, عدّ ونا Our enemy, العقو.بات The sanctions, العقو_بات The sanctions, عنـ ف Violence, عنصر،ي Racist, غ•ز.ه Gaza, غZ ـة Gaza, غزة Gaza, غـزو غزـة Invasion of Gaza, غغز، ة Gaza, فا.رغ. Empty, فا_شل Failure, الفا_شل The failure, الفاسـ_دة The corrupt, فاشل Failure, فاضـ،حة Scandalous, فلسـ طين Palestine, الفلس_طينية The Palestinian, فلسإطين Palestine, فلس، طين Palestine, الفو،لاذ The steel, القـ_تال The fighting, ق،،تل Killing, قو_اعد Bases, القو_اعد The bases, القو9اعد The bases, القوءاعد The bases, كا_يوس Nightmare, كا_ذب Liar, كار،ثه Catastrophe, الكذـ_ب The lying, كذْبي،ة Lie, كمـ ـ ين Ambush, الكنقرس The Congress, كو شنر Kushner, كيء Cauterization, لـ ايـ ران To Iran, لـوحيم Flesh, لأجند_اتكم For your agendas, لإسـ_قاط To overthrow, لاسراااااااااائ ي ل To Israel, لامريييك اااا To America, لأير.ان To Iran, لايرا•ن To Iran, لإيرّان To Iran, لبلطجته For his thuggery, لضر.ـ_بنا To strike us, اللعـ،ين The cursed, لعن/ـة Curse, لغز_ة For Gaza, لكار/ثة For a catastrophe, للطائرّات For the airplanes, للقصف،ف For the bombardment, لنسـ فت To sweep away, لنس فت To sweep away, لهجو_م To sweep away, الم.وت For an attack, المـ،وت The death, المتص هـ ـينين. The Zionist-sympathizers, المتط رفة The radical, المتهو×رة The reckless, المج،،اهد The Mujahid fighter, مجZره Massacre, مَحـ،رَقة Holocaust, محا.كمته ويكر.ههم His trial and hates them, المخـ،تل The flawed/occupier, المخصـ.ب The enriched, مرضـ.نا Our illness, مرّعوب Terrified, المسـ.فوح The slaughtered, المشـ'اهد The scenes, المعـ.ركة The battle, المع،توه The idiot, مهـ_جمة Attacking, مهرـ_جين Clown, مهرـ_جين Clowns, المهر، جين The clowns, ن99ي Nuclear, الن99ي The nuclear, النا.ر The fire, نتـ n يا ه Netanyahu, النتن مات ياهووو The rotten one died yahoo, النتن.يا.هو Netanyahu, نتنـ_ياهو Netanyahu, نتنيا#هو Netanyahu, نتنيا.هو Netanyahu, نتنيا_هو Netanyahu, نتنياهو.. هو Netanyahu, نتيا..هو Netanyahu, نتيا.هو Netanyahu, نست،هدف Targeting, نسـ،طها We shoot it down, النسى ـاء The women, النو.. وي The nuclear, النو..نوية The nuclear, النو.وـية The nuclear, نو.وي Nuclear, النو.وية The nuclear, النـو_وية The nuclear, النّو،وي The nuclear, النو'وي The nuclear, نـووي Nuclear, نـووية Nuclear, هـ.جُـ وم Attack, الهـ_زيمة The defeat, اله،جوم The attack, هـ،جوم Attack, هتضـ،،رب Will be hit, هتهاجمهم ضـ،،رب Will attack them / striking, الهج،مات The attacks, هُجـ،وم Attack, هجو .. م Attack, هجو،ما An attack, هزّق Humiliation/Mockery, هيما.رس Practices, و حيّـ.فا And Haifa, وإسر$ائيل And Israel, وإسر،ائيل And Israel, وإسرا،ئيل And Israel, والأذ.ي And the harm, والأمـ.ريكي And the American, والد،مار And the destruction, والنوءو.ي And the nuclear, والولاي.ات And the United States, وترااامب And Trump, وتسـ_تنزف And draining/exhausting, وتهـ،دد And threatens, وحيـ،،فا And Haifa, ودم*رتها And destroyed it, وسفا_لته And his baseness, وسلا،حهم And their weapon, وسنفتقد.هم And we will miss them, وصوار.. يخنا المد.. مرة And our destructive missiles, وضحا،يا And victims, وغطر+ستهم And their arrogance, وفلسـ•طين And Palestine, وقا_حته And his insolence, الولاءيات The United States, ولعـ،ت.. Set on fire, ومح.... وت And erased/wiped out, ومغتصـ بون And rapists, ونتن، ياهو And Netanyahu, وهـ،روب And fleeing, ويتـ،لذذ And enjoys/relishes, ويخل..صوا And they finish, ويكرّههم And hates them, اليـ&ود The Jews, اليـ@ود The Jews, اليـ&ود The Jews, يفجـورها Explodes it, يق،، Verily/Killing, يقتـ.. له Kills him, يه،،ين Insults, يهزّق Humiliates/mocks, اليور.انيوم The Uranium, يبيع،،يط Crying

**Appendix 3: Sample Images, Texts, Translation by Each Model, # of Models Rendering Correct Translation**

| Images with distorted words | Arabic text in images | #Models with correct translations | Claude | Copilot | Gemini | Qwen |
|---|---|---|---|---|---|---|
|  | خطبة الجمعة .. للشيخ/ صالح الحواني .. بعنوان : غZة تموت جوعا | All correct | Friday sermon by Sheikh/Salih al-Hawani — titled: "**Gaza** Dying of Hunger." | Friday sermon – Sheikh Saleh Al-Hawashi… titled: Gaza is dying of hunger. | Friday Khutbah (sermon) .. by Sheikh / Saleh Al-Hawani .. titled: Gaza is dying of hunger. | Friday sermon.. for Sheikh / Saleh Al-Hawani.. titled: Gazans die of hunger. |
|  | نصرا من الله وفتح قريب الموت لامرييييك !!!! الموت لاسرااااااائ ي ل | All correct | Victory from God is near. Death to America!! Death to Israel!! | Clocked "Victory from God and a near conquest — death to Ameriiiicaaaa, death to Israeeeeeeeel." | Victory from God and an imminent conquest. Death to America !!!! Death to Israiiiiiiil | Victory from God and a close conquest. Death to America!!!! Death to Israel!!!!!!! Y Y L. |
|  | صـواريخ إيـران تهز تـل أبـيب.. قراءة في حجم الخسائر الإسرائـيلية... | All correct | Iranian missiles shake Tel Aviv… a reading on the scale of Israeli losses… | Iran's missiles shake Tel Aviv… a reading of the size of Israeli losses… | Iranian missiles shake Tel Aviv.. A reading of the scale of Israeli losses… | Iranian missiles shake Tel Aviv.. reading the size of Israeli losses… |
|  | السعودية تقلب الطاولة فلس، طين. المنطقة لم تعد تحتمل حرباً جديدة. أنتم من بدأ الحرب وعليكم تحمل أوزارها. إس ، رائيل سبب في حروب المنطقة والفوضى منذ احه، تلالها فك ارتباط النفط بالدولار إلغاء الاتفاقيات والتفاهمات مع الولايات المتحدة تلوح في الأفق ..... | All faulty | Saudi Arabia overturns the table: enough, **mud** [**a metaphor for failure/mess**]. The region cannot withstand a new war. You started the war and you bear it. Your weapons, Israel, are the reason for wars in the region and **chaos** since the link between oil and the dollar was severed, and the cancellation of agreements and understandings with the United States are looming on the horizon. | Saudi Arabia flips the table. Enough—**mud and mess**. The region can no longer bear a new war. You are the ones who started the war, and you must bear its burdens. Israel has been the cause of the region's wars and chaos **since its beginning.** Following that came the link between oil and the dollar… The cancellation of agreements and understandings with the United States is looming on the horizon……" | Saudi Arabia turns the tables on Palestine. The region can no longer tolerate a new war. You are the ones who started the war and you must bear its burdens**. Isr.ael** is the cause of the wars and chaos in the region since its **inception**, forget your connection of oil to the dollar. Cancellation of agreements and understandings with the United States looms on the horizon…… | 13. Saudi Arabia turns the table: **Philistine**, **Hattin**, the region can no longer tolerate a new war. You are the one who started the war and you must bear its burdens. **I.S.rael** is the cause of the region's wars and chaos **since its existence**. For destruction, linking oil to the dollar, canceling agreements and understandings with the United States is looming on the horizon…… |

| | | | | | | |
|---|---|---|---|---|---|---|
|  | عاجل: وزير الخارجية القطري لولوة الخاطر يخاطب ترامب ونتن، ياهو قائلاً : "كُفّوا عن التحدث باسمنا .كفّوا عن استخدامنا كذريعة لأجند_اتكم. لا نريد منكم "تحر_يرنا"، كل ما نريده هو أن تتركونا وشأننا . كفّوا عن تأجيج الحر_وب. ليس ذنبنا أنكم فشلتم في الدراسة وتلقيتم تعليمكم على يد هوليوود، فالعالم ليس فيلماً". | Only one | Urgent — The Qatari Foreign Minister addresses Trump and the women, saying angrily: "Stop talking about us." "Stop using deceitful methods; we don't want your military bases, we don't want anything you deceptions. offer and your "Stop the warmongering; we are tired of it. You've failed in studying and have not learned, although you are Hollywood, not the world, like a movie." | Urgent: Qatari Foreign Minister Lolwah Al-Khater addresses Trump and Netanyahu: “Stop speaking in our name. Stop using us as a pretext for your agenda in the Middle East. Everything you want —we do not want. Your problems have harmed the peoples of the region. You failed in politics and education — your **Harvard** degrees mean nothing in the real world.” | Urgent: Qatari official Lulwah Al-Khater addresses Trump and Netanyahu, saying: "Stop speaking in our name. Stop using us as an excuse for your agendas. We do not want you to 'liberate' us; all we want is for you to leave us alone. Stop fueling wars. It is not our fault that you failed in your studies and received your education at the hands of Hollywood—the world is not a movie." | Minister: The first British Foreign Secretary... saying: "Stop talking in names. Stop using our... Stop... 'went angry', stop... is that you cursed and thanked. Stop the welcome. Sil Anidh... I found in the Sardar and you enjoyed your education on... and Holoho, the world will be my hope." |
|  | 139: ما يزيد احتمالات التعبئة **المتط رفة** ويغذي موجات **عنـ ف** جديدة | 2 correct | The chances of the extremist [**cell/group**] increase, fueled by new waves [**redacted**]. | 138. What increases the chances of the **samp...led** group and enriches the waves of a new **att...ack.** | What increases the chances of radical mobilization and fuels new waves of violence. | What increases probabilities The tense atmosphere feeds New waves of... |
|  | 136: هؤلاء الأسRى داخل سجWن يا الله يا الله 🥺 | 2 correct | These [**people**] are inside an **R**sian prison** 😱 Oh God, oh God [**redacted**]. | These prisoners inside the S prison... oh God, oh God 😢 | These prisoners are inside a prison, Oh God, Oh God 😭 | These are the **US Rs** inside **SJWn** Oh God, oh God |
|  | "النتن مات ياهووو؟! 😱 الصواريخ دخلت المكتب وهو فص ملح وداب.. الحقيقة الكاملة! | 2 correct | "Mom, did you die?! The missiles entered the office, and it was **a total loss...** the complete story!" | The tobacco man is dead, yoooo!? 😳 The missiles entered the office and he vanished like salt in water... the full truth! | "The filthy one [**Netanyahu**] died, yahoo?! 😱 The missiles entered the office and he dissolved like salt in water.. The whole truth! | "Did **Netanyahu** die? Missiles entered the office and it's a piece of salt that dissolved.. The full truth! |
|  | بعد كذْب،ة تدْمـير النّو،وي هآرتس تفجـر مفاجأة إسـرائيل وأمـ... | 3 correct | After the lie... the missile was destroyed... Haaretz reveals: Israel and America... | 129. After the lie of destroying the nuclear program, Haaretz explodes a surprise for Israel and Amer... | After the lie of destroying the nuclear [program], Haaretz drops a bombshell surprise: Israel and **Am[erica]...** | After the **destruction of the West,** yesterday saw the **devastation** of **Israel** and Ma... |
|  | عـاجل: أيها المشـاهد، معلومة خطـ! يرة عليك معرفتها الآن! | 3 correct | Urgent: O viewer, important information you need to know now! | 114. Urgent: Viewer, **this information is wrong**! You must know it now! | Urgent: Dear viewer, critical information you must know right now! | Urgent: O viewer, important information you must know now! |