

Synthetic Slang Generation for Benchmarking Machine Translation

Eric Bennett

Northeastern University

bennett.er@northeastern.edu

Abstract

Slang presents a persistent challenge in machine translation (MT) because its informal, socially situated meanings can depend heavily on context, especially when a system has not encountered the form or sense in training. We introduce a language-adaptable pipeline for generating and validating realistic synthetic slang, and use it to create a benchmark of 1,087 items across Chinese, Russian, and Farsi. We evaluate ten large language model (LLM)-based MT systems under three conditions: translation from a single usage, translation with an additional example of usage in context, and translation conditioned on a gold definition. We find additional context consistently improves translation quality, especially on slang types which are most difficult without assistance. The benefit of providing the gold definition generally grows as model size shrinks. We also test the ability for models to recover the synthetic slang definition, rising to 99% accuracy with 2 examples in context. These results demonstrate how synthetic slang can provide a controlled stress test of meaning inference from context in MT systems.

1 Introduction

Slang translation is a uniquely difficult problem in Machine Translation (MT). Slang is constantly changing and highly individualistic – two friends might use specific terms unused anywhere else in the world in that context. Rather than test the ability for an MT system to retrieve existing cross-lingual representations seen in the training data, this type of language requires an ability to extract the meaning of a term from its context alone.

In this paper, we introduce a method for generating synthetic-but-realistic slang to benchmark MT systems¹, and evaluate it on ten LLM-based MT systems for Chinese, Russian, and Farsi. Using

¹We release code at the following link: <https://github.com/softly-undefined/synthslang>

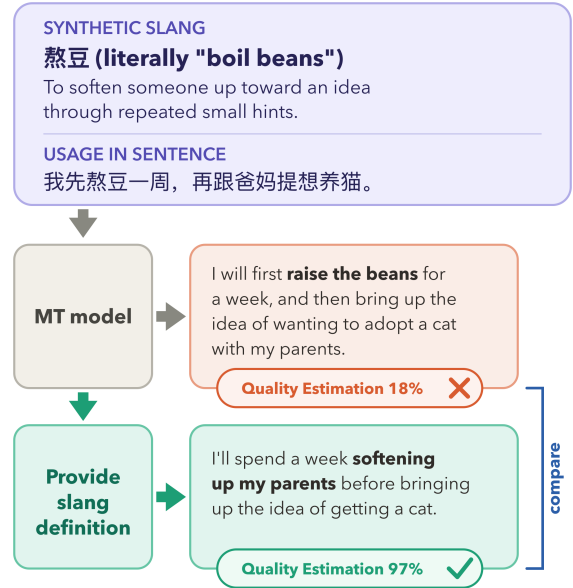


Figure 1: Illustration of the synthetic-slang translation conditions. Given a source sentence containing a synthetic slang term, an MT system translates the sentence with and without access to the term’s intended meaning. Translation quality is compared using reference-free quality-estimation metrics.

the benchmark we report results comparing direct translations of synthetic slang in context to translations conditioned on the ground-truth slang definition, as shown in Figure 1.

1.1 Why Slang is Hard to Translate

Slang and informal language show up constantly in realistic material, and standard MT systems and static lexicons often can’t keep up with how quickly that vocabulary shifts. A term can fall out of use, narrow, broaden, or pick up a completely new meaning within a matter of months, which means a glossary or MT system trained a year ago may already be missing the terms that matter most in current events. Culture-specific references add another layer of complexity, introduc-

ing idioms, running jokes, references to current events, regional slang, and in-group shorthand. A word that reads as neutral or even funny in one cultural context can land very differently in another, and there’s no way to fix that with lexical accuracy alone; the missing piece is background knowledge, not vocabulary.

The difficulty of translating slang differs depending on the linguistic makeup of a language. In Russian, derivational morphology often contributes meaning beyond simple lexical modification. Affixes may encode evaluation, emotional stance, intensity, familiarity, irony, or informality alongside their grammatical functions. Within a constructionist framework, speakers can interpret novel slang formations because they possess abstract knowledge of productive morphological schemas and the semantic and pragmatic meanings associated with them.

Chinese slang is highly dependent on compounding, semantic extension, homophony, and character-level wordplay, so meaning is often carried by the combination and reinterpretation of existing characters. A novel slang term may combine familiar characters in an unfamiliar way, or exploit phonetic similarity with another word. In the development of our Chinese slang generation taxonomy, we analyzed the internet slang data collected by [Zhao et al. \(2026\)](#) for recurring slang patterns.

Farsi forms much of its slang through the productive compound (light-verb) construction, pairing a nominal or adjectival element with a small set of light verbs to coin new verbs whose meaning exceeds their parts. It also draws on ready assimilation of foreign loanwords, which are re-inflected as native words.

Sociolinguistic factors also contribute to the creation of slang, as new slang often acquires meanings associated with particular communities, identities, generations, or communicative settings. Translation of slang language therefore must encode semantic, morphological, and pragmatic information into a single expression, and artfully reinterpret those layers of meaning in another language.

1.2 Slang in Machine Translation

In this work, slang is characterized by informal, socially situated use, and may involve a new form or a new sense. Our synthetic slang spans multiple categories of how slang is developed, including a recently coined form or meaning now

in broader use (neologisms), and a word created for a particular occasion without broader circulation (nonce words), as well as overlapping with related linguistic phenomena, including informal language used in everyday settings (colloquial language), and alternation between languages within discourse (code-switching).

Recent benchmarks have begun to evaluate slang in machine translation. SlangDIT tests whether LLMs can infer and translate the meanings of attested slang expressions from context ([Liang et al., 2025](#)), and [Zhao et al. \(2026\)](#) evaluates translation of Chinese social-media language that includes naturally occurring internet slang. These resources establish slang interpretation as a meaningful stress test of MT. NEO-BENCH studies the effect of neologisms on LLM robustness, including machine translation, using terms that emerged after the models’ knowledge cutoffs ([Zheng et al., 2024](#)). This reduces the likelihood that models can rely on encountering test data during pretraining, but defining novelty relative to cutoff dates ties the benchmark to particular models and time periods, and requires new items as language evolves. We pursue the same goal by using realistic but nonexistent slang, creating a fixed benchmark whose novelty does not depend on when a model was trained.

Synthetic words have been used to test whether a model can apply learned linguistic patterns to unfamiliar forms, taking inspiration from Wug tests, where a learner is asked to apply familiar morphological rules to invented words ([Berko, 1958](#)). [Anh et al. \(2024\)](#) apply a multilingual Wug Test across six languages and find that LLMs can generalize to novel words, though performance varies with morphological complexity. [Pantelidou et al. \(2026\)](#) use a related test and find that language-community size predicts model accuracy better than grammatical complexity. [Ismayilzada et al. \(2025\)](#) test novel roots in Turkish and Finnish and find that models struggle as morphological complexity increases.

Prior work has also shown that MT systems can benefit from external lexical information supplied at inference time. The WMT 2023 Terminology Shared Task directly compares translation with and without segment-level terminology dictionaries ([Semenov et al., 2023](#)). Recent work extends this idea by adding retrieved terminology or broader linguistic descriptions to the prompt ([Zhang et al., 2024](#); [Li et al., 2025](#)).

In this work, we bridge the gap between glossary-assisted translation and synthetic language evaluation by generating synthetic slang that remains linguistically and culturally grounded in each source language, using it to test how MT systems translate novel slang meanings and whether added information about a term’s meaning at inference time helps preserve it in translation.

2 Methodology

Our pipeline has three stages: slang generation, translation, and translation evaluation. We first generate candidate slang terms and filter them through a series of quality and contamination gates; we then translate the surviving items under different context and definition conditions; finally, we evaluate translation quality with validated reference-free metrics and measure definition recovery.

2.1 Slang Generation

We use an LLM to generate potential new slang terms, as well as the usage of such term in both a sentence and paragraph setting. Each term is passed through a series of gates to ensure the quality of the proposed term in order to address the potential flaws with LLM generation of the novel slang terms.

Slang is linguistically and culturally grounded, so prompts are individualized per-language. We use a single shared slang type across all languages (Semantic Shift), and propose 2-3 language-specific slang types. Examples of each type can be found in Appendix A.

2.1.1 Shared Types

Semantic Shift Take an ordinary word in the source language and assign it a completely new figurative meaning through metaphor or metonymy, unrelated to (or only obliquely connected with) its dictionary definition.

Morphological Blend (Shared across Chinese and Russian) Fuse two existing roots, or characters, into a single new word, whose meaning goes beyond either component.

Affixation (Shared across Chinese and Farsi) Attach a prefix or suffix to a base to coin a term for a behavior or social group. Focuses on a small group of affixes, individualized by language.

2.1.2 Chinese Slang Taxonomy

Sentence Compression A process that forms a novel *chengyu*, a classical four-character idiom. Takes a multi-clause colloquial expression and extracts one key character from each clause to produce a memorable 4-character idiom.

2.1.3 Russian Slang Taxonomy

Suffixation Attach a productive colloquial Russian suffix to an existing or invented root to coin an expressive new noun, verb, or adjective for a state, person-type, or behavior.

Foreign Verb Take a non-Russian word (English, French, German, etc.) and russify it by adding native verbal or nominal morphology (-*ить*, -*овать*, -*нуть*, -*уровать*, or noun endings) so it sounds and inflects like a native Russian word.

2.1.4 Farsi Slang Taxonomy

Light Verb Pair a noun or adjective (native, or an assimilated English/foreign loanword) with a native Persian light verb to coin a new compound verb whose slang meaning goes beyond the literal parts.

2.2 Contamination Gates

For each slang type and language we generate 400 examples using OpenAI’s GPT-5.4, requesting 20 slang–definition pairs per call². These examples are then passed through a series of contamination gates to ensure the synthetic slang passed to the translation phase is novel and realistic. We set 100 cleaned items per type as a target, so generating 400 examples per category allows us to apply conservative filtering at each gate. The gates run cheapest-first, so that the more expensive checks only run on surviving candidates, and only terms that clear all four are taken forward. The purpose of the gates is to optimize which segments are shown to human reviewers by removing clearly unsuitable segments. Throughout, we write $\phi(\cdot)$ to represent a fixed multilingual sentence encoder³ that returns an L_2 -normalized embedding, so that the cosine similarity of two texts is their inner product.

Gate 1: Deduplication Because generation only removes string duplicates of slang term, two items may still refer to the same concept. We remove

²The generation prompts are provided in Appendix C

³sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

these near-duplicates by embedding each candidate’s English definition: a candidate is dropped as a near-duplicate of an earlier representative when its cosine similarity to that representative exceeds a cutoff

$$\tau_{\text{dup}} = \max(f, \mu + k\sigma), \quad (1)$$

where μ and σ are the mean and standard deviation of all pairwise similarities in the batch, f is a floor, and k scales the outlier (we use $f=0.5$, $k=3$). The floor ensures that a diverse batch flags nothing, while the relative term adapts the cutoff to each batch.

Gate 2: Guessability A term is only useful to test if its meaning cannot be recovered from its surface form alone. Using the same model that generated the item, we generate n independent zero-context guesses $g_1, \dots, g_n$ of the meaning, showing the term (and, for Russian, its part of speech), and withholding the definition d . We score guessability by the mean cosine similarity between each guess and the true definition,

$$G(t) = \frac{1}{n} \sum_{i=1}^n \cos(\phi(d), \phi(g_i)), \quad (2)$$

and discard the term when $G(t) \geq \tau_g$ (we use $n=3$, and $\tau_g=0.55$). A high score means the form gives the meaning away, and would not be genuinely challenging to translate. The guessability prompt can be found in Appendix C.2.

Gate 3: Existence Check This gate removes terms that already exist with their synthetic slang meaning, a common issue in early outputs. Each term is searched by itself, in quotation marks, and through queries that combine it with information from its definition. Each retrieved snippet is then evaluated in two ways. The embedding judge counts a snippet h as using the intended meaning when $\cos(\phi(d), \phi(h)) \geq \tau_s$. GPT-5.4-mini separately checks whether the snippet uses the intended meaning and flags cases where the term already exists as another word or name. Letting $S(t)$ be the number of distinct domains contributing an in-sense snippet, we remove the term when

$$S(t) \geq N, \quad (3)$$

treating distinct domains as independent attestations (we use $N=3$). In the consensus configuration both judges run off a single retrieval: a term is

Lang.	Slang Type	Gen.	Remaining at gate			Human
			Dedup	Guess	Exist	
Chinese	Semantic Shift	400	324	288	281	100 (57%)
	Morphological Blend	400	322	224	213	100 (65%)
	Affixation	400	294	123	114	87 (76%)
	Sentence Compression	400	248	135	132	100 (77%)
Russian	Semantic Shift	400	329	317	301	100 (60%)
	Morphological Blend	400	334	244	214	100 (76%)
	Suffixation	400	307	276	258	100 (66%)
	Foreign Verb	400	326	275	256	100 (59%)
Farsi	Semantic Shift	400	331	329	324	100 (90%)
	Affixation	400	321	149	146	100 (89%)
	Light Verb	400	329	277	270	100 (89%)

Table 1: Number of synthetic slang items remaining after each contamination gate, by language and slang type. The Human column reports accepted items, with the percentage of reviewed items accepted in parentheses, rounded to the nearest whole percent.

removed only when both flag existence, and cases where the judges disagree are escalated to human review.

Gate 4: Human Validation Terms that survive the first three gates then have a sentence, and paragraph, of the synthetic slang in context generated by an LLM. The resultant (slang, definition, sentence, paragraph) sets are validated by a speaker of the source language. Annotators were asked to measure two aspects of the synthetic slang with a binary judgement: 1) *Faithful Use*: Does the slang term, as used in the sentence and paragraph, faithfully reflect the ground truth definition? 2) *Existence Check*: Is the slang term, as it is defined in definition, novel in this language?

If either aspect is judged as false, we remove that item from the final set. For each slang type we stop human annotation after validating 100 items of that type⁴. Exact details on how many slang items remained after each gate are found in Table 1.

2.3 Slang Translation

We evaluate ten LLM-based MT systems, varying the level of context provided to observe its impact on the translation quality of synthetic slang. All systems used temperature 0 for decoding. The ten evaluated systems are:

- Gemma-3 4B-it, 12B-it, and 27B-it
- Qwen3.5 4B, 9B, and 27B
- GPT-4o-mini
- DeepSeek-V3.2
- Claude Sonnet 4.6
- Gemini 3.5 Flash

⁴For the Chinese Affixation slang type after all validation only 87 terms remained

Reasoning was enabled for Claude Sonnet 4.6, where it was set to high, and Gemini 3.5 Flash, where it was set to medium. Through translation, we examine (1) translation quality and (2) definition recovery. All prompts are provided in Appendix C.3.

Translation Paths

- **Direct:** A simple translation prompt from the source language to the target language. Run on both a sentence and a paragraph. The prompt is based on Andrew Ng’s translation agent⁵.
- **DirectContext:** This prompt only runs on the sentence level. The same prompt as Direct, but the prompt includes the paragraph-level text, without a translation of the text, with a second example of the usage of a term in context.
- **DirectWithHint:** In this prompt method, the model is tasked with translating at both the sentence and paragraph level, conditioned on the slang term and ground-truth definition generated in the Slang Generation phase. This condition mimics that of a retrieval-augmented LLM-based MT system, which has recently become a popular idea.

Definition-Recovery Paths

- **SlangDITIdentify:** Given the slang term and its usage in a sentence, the model is prompted to create a definition of the term. The prompt is inspired from the vanilla prompt in Liang et al. (2025).
- **SlangDITIdentifyContext:** Given the slang term, its usage in a sentence, and its usage in a paragraph, the model is prompted to create a definition of the term.

2.4 Slang Translation Evaluation

With the absence of a true ground-truth translation of these nonexistent slang terms in context, we rely on Quality Estimation (QE) style metrics which take in a (source, hypothesis) tuple and output a score. We do not synthesize reference translations as such translations would be structurally similar to the DirectWithHint translation path, and would themselves require further human validation. QE

metrics require no ground-truth reference, making them well suited to our setting where no such reference exists. However, we acknowledge that these metrics are themselves imperfect proxies for translation quality, and use a human evaluation to select which metrics to report. We consider the following metrics:

1. **CometKiwi:** A reference-free neural QE model (wmt22-cometkiwi-da) that concatenates source and translation and feeds them to a pretrained multilingual encoder (XLM-R/InfoXLM) to regress a single scalar quality score. (Rei et al., 2022)
2. **GEMBA-DA (Direct Assessment):** An LLM-as-a-judge metric that, given a source and hypothesis translation, produces a score on a 0-100 scale. We use OpenAI’s GPT-5.4-mini for scoring. (Kocmi and Federmann, 2023)
3. **GEMBA-DA-Def (ours):** The GEMBA-DA judge conditioned on the synthetic slang term and its gold definition, which is added to the prompt so the model scores against the intended meaning of the novel slang term.

We measure definition-recovery accuracy using a binary LLM judgment of whether the definition generated by SlangDITIdentify or SlangDITIdentifyContext conveys the same core meaning as the gold definition. This metric evaluates semantic equivalence rather than exact wording. Appendix C.4 gives prompts for GEMBA-DA-Def and the binary definition-equivalence judge.

Metric Validation via Human Judgment Selecting a metric is unusually difficult in our setting: because every slang term is synthetic, no existing metric has observed these terms, and learned QE models were trained on standard, naturally occurring language, so their quality signals may not transfer to synthetic slang. We conduct a human evaluation on a subset of translations to measure agreement between a human judge and automatic metrics. A native speaker of Russian and two annotators fluent in Chinese and Farsi, respectively, performed a blinded three-way preference evaluation: for each slang term (sentence-level, Gemma-3 4B outputs), the annotator was shown the Direct and DirectWithHint translations in randomized order and selected the better translation, or

⁵<https://github.com/andrewyng/translation-agent>

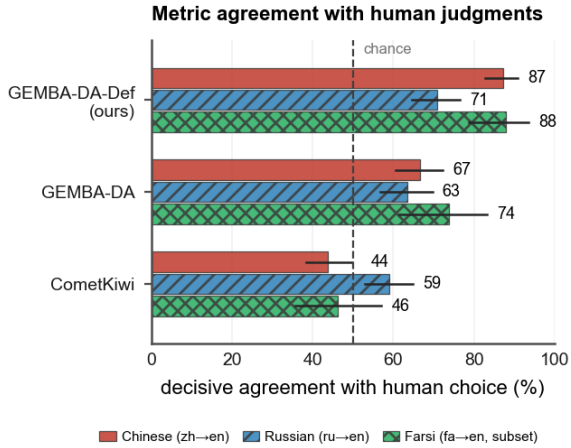


Figure 2: Metric agreement with blinded human judgments across Chinese, Russian, and Farsi. GEMBA-DA conditioned on the gold slang definition (GEMBA-DA-Def) aligns most closely with human preference; Farsi reviews half of all segments, selected randomly.

a tie. We score each candidate metric by its concordance with these judgments, the rate at which the metric’s preferred translation matches the annotator’s choice, reported in Figure 2. Among the metrics tested, GEMBA-DA-Def achieved the highest decisive agreement with human judgments in all three languages: 87.3% for Chinese, 71.0% for Russian, and 88.0% for Farsi. In the Results section we report GEMBA-DA-Def and GEMBA-DA; Appendix B gives a further comparison with two stronger judge variants.

3 Results

A full table of translation-quality results across LLM-based MT systems and metrics can be found in Table 2; definition-recovery results across the underlying LLMs can be found in Table 3.

3.1 Effect of LLM Size

We report results across Gemma-3 (4B, 12B, 27B) and Qwen3.5 (4B, 9B, 27B) size ladders. Figure 3 shows as LLM size increases, translation quality increases, across all definition conditions. However, as LLM size increases, the benefit of providing the gold definition generally decreases. This suggests larger LLMs are better able to interpret the novel slang’s meaning from context alone, and the added information of the hint, or usage in a paragraph, doesn’t make as much of an impact. Notably, across model families, a 4-billion-parameter LLM given the gold definition matches or outperforms a 27-billion-parameter LLM with-

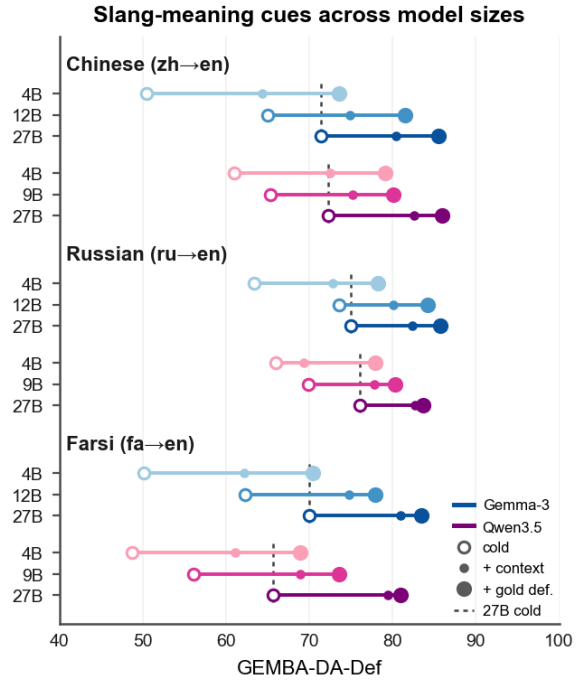


Figure 3: Slang-meaning cues across LLM sizes. Mean GEMBA-DA-Def scores are shown for the Gemma-3 (4B, 12B, and 27B) and Qwen3.5 (4B, 9B, and 27B) size ladders. Translation quality rises with LLM size across both Gemma-3 and Qwen3.5, but the benefit of gold-definition cues shrinks as LLM size grows.

out it.

3.2 Language Specific Results

Figure 4 shows substantial variation in translation quality across slang types, both within and between languages. The Direct condition, where systems must infer the term’s meaning from the sentence-level use alone, shows the most variation. Both assisted conditions produce narrower differences across slang types.

Scores for Semantic Shift differ across languages despite representing the same formation process. For Farsi, additional context brings performance nearly on par with access to the gold definition (74.8 compared with 76.6). One possible explanation is that the mapping between a word’s original and synthetic meanings is not equally transferable across languages. For example, Chinese word formation frequently draws on character-level composition, homophony, and culturally specific associations, so a novel sense may compete with several existing interpretations.

The slang types with the lowest Direct scores consistently receive the largest proportional gains from additional context. Across the 11 language-

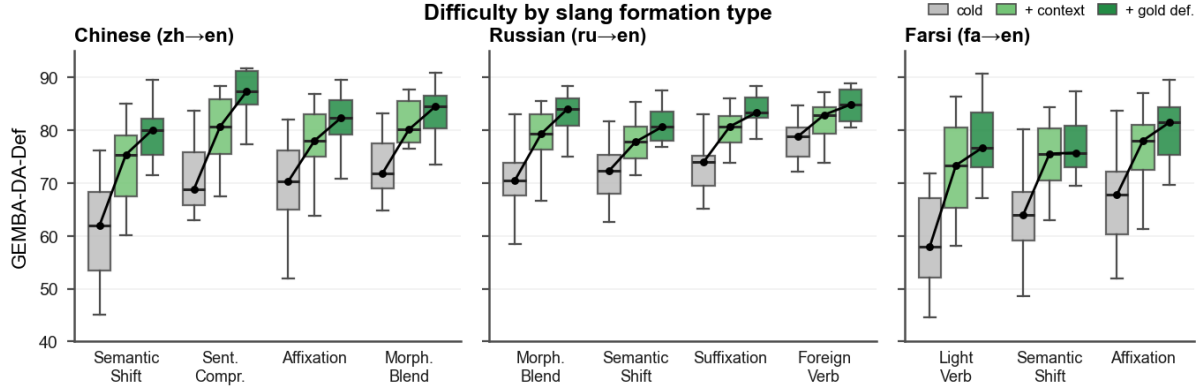


Figure 4: Distribution of per-system mean sentence-level GEMBA-DA-Def scores across ten LLM-based MT systems for the Direct (cold), DirectContext (+context), and DirectWithHint (+gold def.) conditions in Chinese, Russian, and Farsi. Added context produces more consistent translation quality across slang types.

Condition	Lang.	Metric	Gemma-3 4B	Gemma-3 12B	Gemma-3 27B	Qwen3.5 4B	Qwen3.5 9B	Qwen3.5 27B	GPT-4o -mini	DeepSeek -V3.2	Claude Sonnet 4.6	Gemini 3.5 Flash
Direct	Chinese	GEMBA-DA	58.4	72.8	77.6	69.0	73.8	79.2	70.5	79.2	<u>81.5</u>	81.9
		GEMBA-DA-Def	50.5	65.0	71.4	60.9	65.3	72.3	63.0	73.7	<u>77.0</u>	81.3
	Russian	GEMBA-DA	72.5	82.2	84.0	73.0	78.3	83.8	78.6	83.7	<u>84.2</u>	87.1
		GEMBA-DA-Def	63.3	73.6	75.0	66.0	69.9	76.1	71.4	75.3	<u>78.5</u>	83.2
	Farsi	GEMBA-DA	60.4	74.7	80.3	60.5	68.9	77.1	72.9	77.3	<u>82.2</u>	85.6
		GEMBA-DA-Def	50.2	62.3	70.0	48.7	56.1	65.7	60.5	64.8	<u>73.8</u>	78.6
DirectContext	Chinese	GEMBA-DA	65.9	74.1	78.7	70.4	73.6	79.6	74.7	<u>80.7</u>	80.1	82.4
		GEMBA-DA-Def	64.4	74.9	80.4	72.5	75.2	82.5	74.9	82.9	<u>86.0</u>	87.0
	Russian	GEMBA-DA	76.7	82.9	84.5	70.9	80.7	83.9	79.7	82.9	<u>84.9</u>	86.4
		GEMBA-DA-Def	72.8	80.0	82.4	69.3	77.9	82.6	76.7	80.5	<u>84.5</u>	86.0
	Farsi	GEMBA-DA	67.6	78.2	<u>82.3</u>	66.4	72.3	79.3	77.4	78.6	82.1	84.8
		GEMBA-DA-Def	62.2	74.8	81.0	61.1	68.9	79.5	71.3	76.6	<u>84.1</u>	85.9
DirectWithHint	Chinese	GEMBA-DA	65.5	74.8	77.6	71.4	73.4	77.4	72.8	<u>78.1</u>	77.4	81.7
		GEMBA-DA-Def	73.6	81.5	85.6	79.1	80.0	86.0	80.9	85.7	<u>89.0</u>	89.6
	Russian	GEMBA-DA	78.5	83.3	84.9	76.8	79.7	81.8	80.7	82.1	82.9	84.9
		GEMBA-DA-Def	78.3	84.2	85.8	77.9	80.3	83.7	81.5	84.5	<u>86.4</u>	88.3
	Farsi	GEMBA-DA	69.8	77.2	<u>80.8</u>	68.2	72.3	76.8	74.2	73.1	80.7	84.3
		GEMBA-DA-Def	70.4	77.9	83.5	68.9	73.5	80.9	75.3	78.1	<u>85.9</u>	89.2

Table 2: Translation quality across LLM-based MT systems and metrics for Chinese, Russian, and Farsi, by condition (Direct, DirectContext, DirectWithHint), sentence-level. The highest value per metric within each condition block is shown in **bold** and the second-highest underlined. Claude Sonnet 4.6 uses high reasoning, and Gemini 3.5 Flash uses medium reasoning.

type pairs, lower Direct scores coincided with larger proportional gains from added context, suggesting context is especially useful when cold translation is more difficult. Gold definitions also make performance more consistent across slang types. The range between the highest- and lowest-scoring type falls from 11.1 to 7.7 points in Chinese, 7.1 to 3.7 in Russian, and 8.6 to 3.6 in Farsi between Direct and DirectWithHint, reductions of 31%, 48%, and 59%, respectively. Because the performance gap between slang types narrows when the gold definition is provided, the types appear to differ primarily in how difficult their meanings are to infer.

3.3 Definition Recovery

Definition-recovery results across models for the Identify and Context conditions can be found in Table 3. Results are strong across models, demonstrating that LLM-based MT systems can recover the meaning of previously unseen slang terms from only one or two examples of their usage in context. Providing a second usage example in the SlangDITIdentifyContext condition improves definition-recovery accuracy for every system and language, reaching 99% according to our LLM judge for Claude Sonnet 4.6 on both Chinese and Farsi. Across all models and languages, providing a second usage example increased definition-recovery accuracy by 26.3 percentage points (95%

Model	Chinese		Russian		Farsi	
	Id	Ctx	Id	Ctx	Id	Ctx
Gemma-3 4B	0.36	0.66	0.28	0.65	0.31	0.56
Gemma-3 12B	0.57	0.86	0.56	0.86	0.57	0.86
Gemma-3 27B	0.69	0.94	0.58	0.88	0.65	0.92
Qwen3.5 4B	0.60	0.82	0.35	0.70	0.33	0.61
Qwen3.5 9B	0.60	0.85	0.47	0.83	0.44	0.81
Qwen3.5 27B	0.73	0.96	0.64	0.93	0.62	0.91
GPT-4o-mini	0.65	0.91	0.56	0.87	0.57	0.86
DeepSeek-V3.2	0.78	0.96	0.57	0.91	0.66	0.93
Claude Sonnet 4.6	0.85	0.99	0.79	0.96	0.82	0.99
Gemini 3.5 Flash	0.77	0.97	0.77	0.91	0.81	0.96

Table 3: Definition-recovery accuracy across models for Chinese, Russian, and Farsi. A prediction is counted as correct when it conveys the same core meaning as the gold definition, as decided by an LLM judge. **Id** = SlangDITIdentify, **Ctx** = SlangDITIdentifyContext. The highest value per column is shown in **bold**. Claude Sonnet 4.6 uses high reasoning, and Gemini 3.5 Flash uses medium reasoning.

CI [24.9, 27.7]). This improvement was statistically significant for every model–language combination.

4 Conclusion

We introduce a language-adaptable pipeline for generating and validating synthetic slang and use it to benchmark ten LLM-based MT systems across Chinese, Russian, and Farsi. Across systems, providing an additional usage example or gold definition improves translation quality, with gold-definition gains generally largest among smaller models and gains from either form of assistance largest for slang types that are difficult to translate from a single usage. Systems can also recover the meanings of synthetic slang from context, with a second usage example improving accuracy across every system and language combination evaluated. Together, these results show that synthetic slang can provide a controlled way to test how well MT systems infer unfamiliar meanings from context rather than rely on forms and senses encountered during training.

5 Limitations

The benchmark covers three source languages and eleven language–slang type combinations, all translated into English. These slang types are not exhaustive and cannot capture the full complexity and cultural context of how slang develops in the real world. Future work should compare results on

synthetic slang to recently-attested naturally occurring slang.

To keep translation evaluation annotation manageable, metric validation focused on sentence-level translations from Gemma-3 4B, with only a random half of the Farsi pairs evaluated. These results provide evidence for selecting GEMBA-DA-Def in the tested setting, although broader annotation would be useful to confirm that its agreement with human judgments holds across annotators, models, and source length.

6 Acknowledgments

This work was developed during the Summer Conference on Applied Data Science (SCADS), hosted by the Laboratory for Analytic Sciences at North Carolina State University. Thanks to the language professionals who contributed to human validation and annotation, especially Larysa T.

References

- Dang Anh, Limor Raviv, and Lukas Galke. 2024. [Morphology matters: Probing the cross-linguistic morphological generalization abilities of large language models through a wug test](#). In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 177–188, Bangkok, Thailand. Association for Computational Linguistics.
- Jean Berko. 1958. [The child’s learning of english morphology](#). *WORD*, 14(2-3):150–177.
- Mete Ismayilzada, Defne Circi, Jonne Sälevä, Hale Sirin, Abdullatif Köksal, Bhuwan Dhingra, Antoine Bosselut, Duygu Ataman, and Lonneke Van Der Plas. 2025. [Evaluating morphological compositional generalization in large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1270–1305, Albuquerque, New Mexico. Association for Computational Linguistics.
- Tom Kocmi and Christian Federmann. 2023. [Large language models are state-of-the-art evaluators of translation quality](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.
- Bryan Li, Jiaming Luo, Eleftheria Briakou, and Colin Cherry. 2025. [Leveraging domain knowledge at inference time for LLM translation: Retrieval versus generation](#). In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pages 91–106, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Yunlong Liang, Fandong Meng, Jiaan Wang, and Jie Zhou. 2025. [Slangdit: Benchmarking llms in interpretative slang translation](#). *Preprint*, arXiv:2505.14181.

Nikoleta Pantelidou, Evelina Leivada, Raquel Montero, and Paolo Morosi. 2026. [Community size rather than grammatical complexity better predicts large language model accuracy in a novel wug test](#). *Preprint*, arXiv:2510.12463.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. [CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. 2023. [Findings of the WMT 2023 shared task on machine translation with terminologies](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 663–671, Singapore. Association for Computational Linguistics.

Kexun Zhang, Yee Man Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. [Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15654–15669, Bangkok, Thailand. Association for Computational Linguistics.

Kaiyan Zhao, Zheyong Xie, Zhongtao Miao, Xinze Lyu, Yao Hu, and Shaosheng Cao. 2026. [Benchmarking machine translation on chinese social media texts](#). *Preprint*, arXiv:2601.22931.

Jonathan Zheng, Alan Ritter, and Wei Xu. 2024. [NEO-BENCH: Evaluating robustness of large language models with neologisms](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13885–13906, Bangkok, Thailand. Association for Computational Linguistics.

A Slang Examples

Table 4 contains representative examples of synthetic slang generated for each language and slang type.

B Metric Selection Details

To test whether additional compute in the LLM-judge process improves agreement with human judgments, we compare the reported GEMBA-DA-Def against two more expensive variants for

the validated Russian slang. All three configurations score the same 400 blinded Direct versus DirectWithHint pairs produced by Gemma-3 4B at the sentence-level. GEMBA-DA-Def-Big uses GPT-5.4 instead of GPT-5.4-mini, while GEMBA-DA-Def-Avg3 samples the GPT-5.4-mini judge three times and averages the resulting scores. Table 5 reports three-way agreement, which treats a match on Tie as agreement, and decisive agreement, which excludes pairs where either the annotator or metric selected Tie. GEMBA-DA-Def-Big produces the highest agreement, reaching 73.8% on decisive pairs compared with 71.0% for GEMBA-DA-Def. Averaging three GPT-5.4-mini judgments does not improve concordance. We identify a single GPT-5.4-mini GEMBA-DA-Def judgment as the best cost–performance tradeoff for the main experiments.

C Prompt Templates

C.1 Synthetic Slang Generation

The following are the run-time prompts for the eleven language–formation combinations used in the paper. For space, all few-shot demonstration blocks are omitted throughout this appendix; all other prompt text and formatting are reproduced verbatim.

Chinese: Semantic Shift

Invent fake-but-plausible Chinese internet slang using SEMANTIC SHIFT: take an ordinary Chinese word and assign it a completely new figurative meaning through metaphor or metonymy.

Rules:

- The new meaning must be non-obvious from the characters alone.
- The meaning should name a recognizable modern phenomenon, behavior, feeling, or situation.
- Do not use words already known to have slang meanings.

Generate exactly 20 fake slang terms as a JSON array:

```
[
  {
    "term": "Chinese characters",
    "base": "literal meaning",
    "definition_tar_lang": "the new slang meaning, in English, 1-2 sentences"
  }
]
```

Chinese: Morphological Blend

Invent fake-but-plausible Chinese internet slang using MORPHOLOGICAL BLEND (混成型): fuse two existing words or morphemes into a new compound whose meaning goes beyond either component.

Rules:

- The result should feel like a single new concept, not just a noun phrase.
- The meaning should not be fully predictable from the parts — there should be a conceptual leap.
- Aim for 2–3 character compounds.

	Slang Type	Term	Definition	Usage in Sentence
Chinese	Semantic Shift	熬豆 áo dòu	To spend a long time softening up a hesitant friend, parent, or group chat toward an idea through repeated small hints rather than one direct ask.	我先熬豆一周，再跟爸妈提想养猫。 <i>I'll work on them gradually for a week first, then bring up wanting to adopt a cat to my parents.</i>
	Morphological Blend	拖嗨 tuō hāi	The weirdly energized state where, instead of doing the task, you become intensely active about everything around it: organizing files, researching tools, making playlists. Procrastination that feels productive and exciting.	本来要写方案，结果先把桌面收拾到拖嗨。 <i>I was supposed to write the proposal, but instead I got so dragged-hyped that I ended up tidying my desktop first.</i>
	Affixation	陪删狗 péi shān gǒu	A friend called in to witness and support a dramatic cleanup session, whether digital or emotional. Their role is not to solve the problem, but to say 'yes, delete it' at the right moments.	失恋后找了个陪删狗，聊天记录终于清干净了。 <i>After the breakup, I got a deletion buddy, and finally cleared out the chat history.</i>
	Sentence Compression	雨停带伞 yǔ tíng dài sǎn	Used for mildly annoying timing irony: you prepare carefully for bad weather and then the weather immediately stops being a problem.	早高峰挤地铁还拎把伞，结果一到园区就雨停带伞。 <i>I squeezed onto the subway during rush hour carrying an umbrella, and as soon as I got to the office park, it was a total 'rain-stops-once-you-bring-an-umbrella' moment.</i>
Russian	Semantic Shift	распарить rasparit'	To wear down someone's resistance to an idea by surrounding them with low-stakes examples, jokes, and repeated exposure until they agree almost passively. It is softening by atmosphere rather than argument.	Он меня так распарил мемами про переезд, что я уже смотрю билеты. <i>He softened me up so much with memes about moving that I'm already looking at tickets.</i>
	Morphological Blend	срачегория srachegoriya	A topic bucket online that reliably stops being discussion and becomes ritualized combat, regardless of how harmless the original post looked.	Стоило упомянуть доставку еды, и тред мгновенно уехал в срачегорию. <i>I only had to mention food delivery, and the thread instantly derailed into a flame-war zone.</i>
	Suffixation	жмакня zhmaknya	The sloppy routine of pressing random interface elements, filters, and reaction buttons without much intent, just to produce movement on screen. It is interaction as idle fidgeting.	Пока ждал созвон, устроил жмакню в настройках и случайно включил тёмную тему. <i>While waiting for the call, I did some zhmaknya in the settings and accidentally turned on dark mode.</i>
	Foreign Verb	блендануть blendanut'	To quietly adapt your style, opinions, or wording to match whatever online crowd you are currently in. It suggests social camouflage so smooth that nobody notices the switch.	Зашёл в новый чат по играм и за вечер так бленданул, будто всегда там сидел. <i>I joined a new gaming chat and blended in so well over the evening, as if I'd always been there.</i>
Farsi	Semantic Shift	تخم مرغ tokhm-e morgh	An idea or plan everyone treats with extreme caution because it is promising but absurdly easy to ruin with one careless comment, delay, or public exposure.	فعلاً درباره همکاری جدید چیزی نگو؛ هنوز تخم مرغ و زود می ترکه. <i>For now, don't say anything about the new collaboration; it's still an egg and could crack too soon.</i>
	Affixation	پزخور poz-khor	A person who is unusually vulnerable to other people's flexing and instantly feels small, provoked, or compelled to answer back. One subtle display of money, taste, or access is enough to ruin their peace.	به استوری از سفر دبی دید، دوباره پزخوریش زد بالا و شروع کرد کنایه انداختن. <i>He saw a story from someone's Dubai trip, got triggered by flexing again, and started throwing snide remarks.</i>
	Light Verb	مهتاب بستن mahtâb bas-tan	To give an ordinary moment a soft, cinematic glow in retelling so it sounds deeper, prettier, or more fated than it really was. Often used for romanticizing mediocre nights, weak dates, or basic memories.	قرارمون فقط به قهوه‌ی معمولی بود، ولی تو استوری‌هاش چنان مهتاب بست انگار فیلم نولان دیده‌ایم. <i>Our date was just an ordinary coffee, but in her stories she moonlit it so much it looked like we'd watched a Nolan film.</i>

Table 4: Representative synthetic slang examples for each language and slang type. Native-script terms are shown with a Latin transliteration beneath. The translations below each sentence are example translations generated by GPT-5.4 and are not considered ground truth.

Metric	3-way	κ	Non-tie n	Decisive
GEMBA-DA-Def	49.2%	0.15	217	71.0%
GEMBA-DA-Def-Big	52.2%	0.19	237	73.8%
GEMBA-DA-Def-Avg3	44.2%	0.07	259	67.6%

Table 5: Agreement between GEMBA-DA-Def judge configurations and blinded human judgments on 400 Russian Direct versus DirectWithHint pairs. Three-way agreement includes matching Tie judgments.

- It should name something people recognize but didn't have a word for.

Generate exactly 20 fake slang terms as a JSON array:

```
[
  {
    "term": "Chinese characters",
    "components": "Word A (meaning) + Word B (meaning) → blend logic",
    "definition_tar_lang": "what the compound means, in English, 1-2 sentences"
  }
]
```

Chinese: Affixation

Invent fake-but-plausible Chinese internet slang using AFFIXATION (派生型): attach a productive suffix or prefix to a novel base to coin a new term for a person-type, social group, or behavior.

Affixes (use a mix of all 6 across your 20 examples):

- 族: social tribe defined by a shared situation
- 控: someone obsessively into X
- 狗: self-deprecating in-group label
- 系: a style or aesthetic vibe
- 裸: doing something stripped-down, without backup
- 被: passive/"reportedly" — done to you without consent, often ironic

Rules:

- The coined term must describe something recognizable but not-yet-named.
- The definition should describe a recognizable kind of person, group, or behavior.
- Do not reuse existing well-known slang.

Generate exactly 20 fake slang terms as a JSON array:

```
[
  {
    "term": "Chinese characters",
    "affix": "which affix and its meaning",
    "base": "root and its meaning",
    "definition_tar_lang": "what the term means, in English, 2-3 sentences"
  }
]
```

Chinese: Sentence Compression

Invent fake-but-plausible Chinese "internet chengyu" using SENTENCE COMPRESSION (缩略型): take a multi-clause colloquial expression and extract one key character from each clause to produce a memorable 4-character idiom mimicking the classical chengyu form.

Construction: start with a natural colloquial expression (2-4 clauses, 20-50 chars), extract the most semantically loaded character from each clause, string them together so they read well aloud.

Rules:

- Source expression must be colloquial modern Chinese, not literary.
- The 4-char form must be traceable back to the source.
- The meaning should reflect a recognizable modern phenomenon, behavior, feeling, or situation.

Generate exactly 20 fake internet chengyu as a JSON array:

```
[
  {
    "term": "4 Chinese characters",
    "source": "full source expression",
    "source_gloss": "English translation of source expression",
    "compression_map": "char1 from clause 1 + char2 from clause 2 + ...",
    "definition_tar_lang": "what it means and when you'd use it, in English"
  }
]
```

Russian: Semantic Shift

Invent fake-but-plausible Russian internet slang using SEMANTIC SHIFT: take a real, ordinary Russian word and assign it a radically new colloquial meaning through metaphor or metonymy, unrelated to (or only obliquely connected with) its dictionary definition.

Rules:

- The base must be a real Russian word, but one with NO documented slang meaning.
- The new meaning must not be predictable from the literal meaning alone — there should be a figurative leap, not a paraphrase.
- The meaning should name a recognizable modern phenomenon, behavior, feeling, or situation.
- The word should keep its normal Russian inflection (conjugation/declension) in the new sense.

Generate exactly 20 fake slang terms as a JSON array:

```
[
  {
    "term": "Cyrillic term",
    "translit": "Latin transliteration of the term",
    "pos": "part of speech",
    "base": "literal/dictionary meaning of the word",
    "definition_tar_lang": "the new slang meaning, in English, 1-2 sentences"
  }
]
```

Russian: Morphological Blend

Invent fake-but-plausible Russian internet slang using MORPHOLOGICAL BLEND: fuse two existing roots (Russian + Russian, or Russian + assimilated foreign) into a single new word, portmanteau-style, whose meaning goes beyond either component.

Rules:

- The result should read as a single new word, not a noun phrase or hyphenated pair.
- The meaning should not be fully predictable from the parts — there should be a conceptual leap.
- The blend must be phonologically tight and pronounceable, and must inflect like a normal Russian word.
- It should name something people recognize but didn't have a word for, and must not be an existing word or known slang.

Generate exactly 20 fake slang terms as a JSON array:

```
[
  {
    "term": "Cyrillic term",
    "translit": "Latin transliteration of the term",
    "pos": "part of speech",
    "components": "Root A (meaning) + Root B (meaning) → blend logic",
    "definition_tar_lang": "what the blend means, in English, 1-2 sentences"
  }
]
```

Russian: Suffixation

Invent fake-but-plausible Russian internet slang using SUFFIXATION: attach a productive colloquial Russian suffix to an existing or invented root to coin an expressive new noun, verb, or adjective for a state, person-type, or behavior.

Suffixes (use a mix across your 20 examples):

- яра, -ак / -як, -ок: expressive personal/object nouns
- няк: a chronic state or condition (cf. бардак-like expressive nouns)
- уха: expressive feminine nouns for states or things
- ня, -ёж, -лово: collective/process nouns, often dismissive
- ота: a quality or atmosphere, with an unexpectedly warm or ironic tone

Rules:

- The coined word must describe something recognizable but not-yet-named.
- The suffix's register and the root's meaning must combine naturally — the word should sound like it could be real.
- The result must not be an existing word or known slang, and the meaning must not be transparently obvious from the root alone.

Generate exactly 20 fake slang terms as a JSON array:

```
[
  {
    "term": "Cyrillic term",
    "translit": "Latin transliteration of the term",
    "pos": "part of speech",
    "affix": "which suffix and the flavor it adds",
    "base": "root and its meaning (real or invented)",
    "definition_tar_lang": "what the term means, in English, 1-2 sentences"
  }
]
```

Russian: Foreign Word Russification

Invent fake-but-plausible Russian internet slang using FOREIGN WORD RUSSIFICATION: take a non-Russian word (English, French, German, etc.) and russify it by adding native verbal or nominal morphology (-ить, -овать, -нуть, -ануть, -ировать, or noun endings) so it sounds and inflects like a native Russian word.

Rules:

- The foreign source word must NOT already be established in Russian slang in this form (avoid well-worn borrowings like чилить, флексить, кринжовать).
- The russified form must inflect naturally (conjugate/decline like a native word) and be phonologically comfortable in Russian.
- Avoid forms that sit close to existing Russian words — especially vulgar ones — in sound or spelling.
- The meaning should add a specific twist or scenario beyond the plain meaning of the source word — not just a translation of it.
- The meaning should reflect a recognizable modern phenomenon, behavior, feeling, or situation.

Generate exactly 20 fake slang terms as a JSON array:

```
[
  {
    "term": "Cyrillic term",
    "translit": "Latin transliteration of the term",
    "pos": "part of speech",
    "source_word": "foreign source word and its language",
    "morphology": "which Russian affix/ending was added",
    "definition_tar_lang": "what it means, in English, 1-2 sentences"
  }
]
```

Farsi: Light Verb Construction

Invent fake-but-plausible Farsi (Persian) slang using LIGHT VERB CONSTRUCTION: pair a noun or adjective (native, or an assimilated English/foreign loanword) with a native Persian light verb to coin a new compound verb whose slang meaning goes beyond the literal parts.

Light verbs to draw from (use a mix across your 20 examples): ندرک (to do/make), نَدش (to become), نذر (to strike), نَتفرگ (to take/catch), نداد (to give), نَدروا (to bring), نَدیَشک (to drag/pull), نَتسب (to tie/close), نَدروخ (to eat).

Rules:

- The compound's meaning must NOT be predictable from the noun/adjective + verb literally — there must be a figurative leap.
- The light verb must be one of the native light verbs above (not an invented verb).
- The meaning should name a recognizable modern phenomenon, behavior, feeling, or situation.
- The compound must conjugate naturally like a normal Persian compound verb.
- Do not reuse existing well-known compound-verb slang (e.g. نَدز غیت, نَدز خِم, نَتسب وِلاخ).

Generate exactly 20 fake slang terms as a JSON array:

```
[
  {
    "term": "Persian term (noun/adjective + light verb)",
    "translit": "Latin transliteration of the term",
    "light_verb": "which light verb and its literal meaning",
    "base": "the noun/adjective and its literal meaning",
    "definition_tar_lang": "the new slang meaning, in English, 1-2 sentences"
  }
]
```

Farsi: Affixation

Invent fake-but-plausible Farsi (Persian) slang using AFFIXATION: attach a productive Persian prefix or suffix to a novel base (native or an assimilated loanword) to coin a new term for a person-type, social group, style, or behavior.

Affixes (use a mix across your 20 examples):

- زاب (-bâz): someone obsessed with, playing at, or generating drama around the root
- روخ (-khor, "eater"): someone who suffers from or greedily consumes the root
- روط (-toor): doing something in the manner / vibe / style of the root
- لام (-mâl, "rubber"): one who flatters or sucks up via the root
- ب (-bi, "without"): a cutting character judgment — lacking the root quality
- هَدنگ (-gondeh-, "huge"): a false, inflated sense of grandeur about the root
- هس (-se-, "three"): doing the root to the utmost / in a flash
- جهل (-chehel-, "forty"): an abundant, chaotic, or excessive amount of the root

Rules:

- The coined term must describe something recognizable but not-yet-named.
- The affix's flavor and the base's meaning must combine naturally — it should sound like it could be real.
- The meaning should describe a recognizable kind of person, group, style, or behavior.
- Do not reuse existing well-known slang (e.g. زابرتخ, هبنجیب, لامه‌ی‌اخ, روخ‌ه‌صغ, زابرتخ).

Generate exactly 20 fake slang terms as a JSON array:

```
[
  {
    "term": "Persian term",
    "translit": "Latin transliteration of the term",
    "affix": "which affix and the flavor it adds",
    "base": "root and its meaning (native or loanword)",
    "definition_tar_lang": "what the term means, in English, 2-3 sentences"
  }
]
```

Farsi: Semantic Shift

Invent fake-but-plausible Farsi (Persian) slang using SEMANTIC SHIFT: take an ordinary, everyday Persian word and assign it a radically new colloquial meaning through metaphor, metonymy, or subversion, unrelated to (or only obliquely connected with) its normal meaning.

Rules:

- The base must be a real, ordinary Persian word with NO documented slang meaning.
- The new meaning must not be predictable from the literal meaning alone — there should be a figurative leap, not a paraphrase.
- The meaning should name a recognizable modern phenomenon, behavior, feeling, or situation.
- The word should keep its normal Persian grammar in the new sense.

Generate exactly 20 fake slang terms as a JSON array:

```
[
  {
    "term": "Persian word",
    "translit": "Latin transliteration of the term",
    "base": "the ordinary/literal meaning of the word",
    "definition_tar_lang": "the new slang meaning, in English, 1-2 sentences"
  }
]
```

After the first batch, the code appends the following block, with one list item for every term already seen in the current or resumed run:

IMPORTANT — the terms below have ALREADY been generated. Do NOT repeat any of them, and do NOT coin new terms that are close to them in form OR in meaning. Every term you return must be clearly distinct from all of these:

- {already_generated_term_1}
- {already_generated_term_2}

C.2 Guessability Gate

You will be shown a {language} slang term with NO context. Guess what it most likely means, reasoning only from its {form_basis} — its form, components, and sound. Do not search; give your single best interpretation.

Term: {term}
{optional_part_of_speech}

Return JSON:

```
{
  "guess": "your best guess at the meaning, in English"
}
```

The value of form_basis is “character components” for Chinese, “morphology and root” for Russian, and “morphology, roots, and component words” for Farsi. The optional line is Part of speech: {pos} for Russian and empty otherwise.

C.3 Translation and Definition-Recovery Inference

For each translation condition below, the first paragraph is the system message and the remainder is the user prompt.

Direct

You are an expert linguist, specializing in translation from {src} to {tgt}.

This is an {src} to {tgt} translation, please provide the {tgt} translation for this text. Do not provide any explanations or text apart from the translation.

{src}: {source}

{tgt}:

DirectWithHint

You are an expert linguist, specializing in translation from {src} to {tgt}.

This is an {src} to {tgt} translation, please provide the {tgt} translation for this text. Do not provide any explanations or text apart from the translation.

{src}: {source}

Note: The text contains the following slang: "{term}" — {definition}

{tgt}:

DirectContext

You are an expert linguist, specializing in translation from {src} to {tgt}.

This is an {src} to {tgt} translation, please provide the {tgt} translation for this text. Do not provide any explanations or text apart from the translation.

For reference, here is a {src} paragraph using similar terminology in context (provided only to help you understand the usage — do not translate it):

{context}

{src}: {source}

{tgt}:

SlangDITIdentify

Your task is to analyze the given sentence, which contains a slang term. Follow the steps below to provide your answer.

1. Slang Term:

- The sentence contains the slang term: "{term}".

2. Provide Definition:

- Provide a brief definition, in English, of the slang term's meaning and usage.

3. Output the Results as JSON with the following keys:

- "slang": the slang term (string)
- "definition": brief definition of the slang term, in English (string)

Please analyze the following sentence:

Input Sentence: {source}

Slang term: {term}

SlangDITIdentifyContext

Your task is to analyze the given sentence, which contains a slang term. Follow the steps below to provide your answer.

1. Slang Term:

- The sentence contains the slang term: "{term}".

2. Provide Definition:

- Provide a brief definition, in English, of the slang term's meaning and usage.

3. Output the Results as JSON with the following keys:

- "slang": the slang term (string)
- "definition": brief definition of the slang term, in English (string)

For reference, here is a paragraph that uses the same term in context (provided only to help you understand its usage):
{context}

Please analyze the following sentence:
Input Sentence: {source}
Slang term: {term}

C.4 Project-Specific Evaluation Prompts

GEMBA-DA-Def

Score the following translation from {src} to {tgt} on a continuous scale from 0 to 100, where a score of zero means "no meaning preserved" and score of one hundred means "perfect meaning and grammar".

Note: the source contains the slang term "{term}", which means: {definition}

{src} source: "{source}"
{tgt} translation: "{mt_translation}"

Return JSON:
{
 "score": 75
}

Binary Definition-Equivalence Judge

Term: "{term}"

Definition A: {reference_definition}
Definition B: {mt_understanding}

Do these two definitions convey the same core meaning?
- 1 = yes: the essential meaning matches (allow paraphrase; minor gaps are fine)
- 0 = no: the meaning is wrong, inverted, or a key component is missing

Return JSON:
{
 "score": 0,
 "reason": "one sentence"
}