Data Contamination in Neural Hieroglyphic Translation: A Reproducibility Study

Ammar Toutou[1] **Abdelrahman Harb**[1] **Christine Basta**[2,3]
[1]Computer Science and Engineering, Alamein International University (AIU), Egypt
[2]HiTZ Center, University of the Basque Country, Spain
[3]Faculty of Computers and Data Science, Alexandria University, Egypt
{ammar.mohamed.2023, abdelraheman.shehata.2023}@aiu.edu.eg
christine.basta@alexu.edu.eg

Abstract

Ancient and endangered languages pose a unique challenge for NLP: their datasets are inherently scarce, difficult to expand, and built from formulaic corpora—making data-quality issues especially consequential yet rarely audited. Motivated by the need to understand what current NMT can realistically achieve for such languages, we investigate hieroglyphic-to-German translation, where a recent study reported 61.5 BLEU using fine-tuned M2M-100. Our reproduction yields only 37.0 BLEU with the released model. Investigating this gap, we find **32% of test targets appear identically in training** (16/50; 50% under 8-gram overlap at 70% threshold). This contamination inflates scores dramatically: contaminated samples achieve up to 83.8 BLEU / 0.924 COMET-22 versus 30.9–39.2 BLEU / 0.622–0.676 COMET-22 on clean samples across five model configurations spanning two architectures. Document-level decontamination reduces contaminated BLEU by only 4.6 points because 8/16 targets persist via other source documents—target-level deduplication is required. We release a decontaminated 34-sample test set and establish corrected baselines (30.9–39.2 BLEU), providing a realistic assessment of NMT capability for this endangered writing system.

1 Introduction

Ancient and endangered languages—including Ancient Egyptian, Akkadian, and Classical Latin—represent some of the most data-scarce domains in NLP (Sommerschield et al., 2023). Unlike modern languages where parallel corpora can be crowdsourced or web-scraped at scale, ancient-language datasets are inherently limited: they depend on surviving archaeological artifacts, require specialist expertise for annotation, and cannot be expanded on demand. The Thesaurus Linguae Aegyptiae (TLA), the largest digitized resource for Ancient Egyptian, yields only **18,669 usable training pairs** after filtering for samples with both digitized hieroglyphic source and German target—orders of magnitude smaller than modern MT benchmarks. This scarcity makes data-quality issues especially consequential: every contaminated or duplicated sample has outsized impact on evaluation.

Motivated by the need to establish what NMT can realistically achieve for such resource-limited languages, we investigated hieroglyphic-to-German translation. De Cao et al. (2024)[1] reported 61.5 BLEU using M2M-100 (Fan et al., 2021) fine-tuned on TLA data—a score approaching neural ceilings on high-resource benchmarks (Popel et al., 2020). Such performance would be remarkable for an endangered writing system with severely limited data. However, our reproduction using the publicly released model yielded only **37.0 BLEU**, a gap of over 24 points.

This discrepancy led us to examine the dataset itself, where we discovered that **32% of test samples have German target translations appearing identically in the training data**. This contamination—arising naturally from formulaic repetition in ancient corpora (medical instructions, offering formulae, royal epithets)—enables models to achieve high scores through memorization rather than genuine translation. On the 16 contaminated test samples, our best model achieves **83.8 BLEU**; on the 34 clean samples, only **37.3 BLEU**—a 47-point gap. This pattern holds across all five models tested (29–47 point gaps), confirming the problem is dataset-inherent.

These findings have broader implications: ancient corpora universally contain formulaic repetition, making train-test contamination likely under standard random splitting. Our corrected baselines (30.9–39.2 BLEU) provide a realistic assessment—useful for corpus triage but requiring

[1]Model publicly available on HuggingFace.

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 50–57

expert verification—and highlight the need for contamination auditing in all ancient-language NLP.

Contributions.

1. **First contamination audit for ancient-language NMT.** We show that formulaic repetition—pervasive in ancient Egyptian corpora—creates systematic data-leakage risks absent from modern-language benchmarks, and identify 32% exact target overlap in the public test set.

2. **Cross-architecture, cross-metric robustness.** The inflation is dataset-inherent, not model-specific: contaminated samples score 29–47 BLEU and 0.23–0.26 COMET-22 points higher than clean samples across five configurations spanning two architectures (M2M-100/mBART-50).

3. **Graduated decontamination analysis.** Beyond exact matching, we apply character 8-gram overlap thresholding, document-level leakage analysis, and per-item frequency cataloguing to characterize contamination severity.

4. **Corrected baselines and reusable evaluation protocol.** We release a decontaminated test set (34 samples, no target overlap), contamination-detection scripts, and corrected baselines (30.9–39.2 BLEU) as a realistic performance range for future work.

2 Background

Hieroglyphic Translation. Ancient Egyptian hieroglyphics (ca. 3200 BCE–400 CE) use a mixture of logographic and phonetic signs. The TLA provides the largest parallel resource of transliterations and German translations; hieroglyphs are encoded in Gardiner notation for computational processing. De Cao et al. (2024) fine-tuned M2M-100 (Fan et al., 2021) on TLA data, reporting 61.5 BLEU for hieroglyphic-to-German.

Data Contamination. Contamination occurs when evaluation targets leak into training. Magar and Schwartz (2022) and Kocyigit et al. (2025) showed 20–30 BLEU inflation in controlled experiments. Ancient corpora are particularly susceptible: formulaic phrases repeat across many texts, so random splitting distributes identical targets across partitions—**a methodological challenge rather than a researcher oversight**. Reproducibility challenges are endemic in ML (Pineau et al., 2021); to our knowledge, no prior study has audited contamination in ancient-language NMT.

Hieroglyphic NLP. Computational approaches have addressed recognition (Franken and van Gemert, 2013; Barucci et al., 2021), transliteration (Wiesenbach and Riezler, 2019), and translation (De Cao et al., 2024). Chen et al. (2024) present the first multi-task benchmark for ancient logographic systems; unlike LogogramNLP, our work audits contamination and provides splitting protocols.

Contamination Detection Methods. Beyond controlled experiments (Section 2), recent work has revealed contamination in deployed systems: Tan et al. (2026) showed cross-direction contamination in FLORES-200; Abbas et al. (2026) demonstrated obfuscation-hidden contamination in Arabic NLP; Enis and Hopkins (2024) identified test data in closed models' training. Detection methods such as permutation testing (Oren et al., 2023), distribution analysis (Dong et al., 2024), and generalization-gap testing (Dekoninck et al., 2024) would be complementary for future closed-model evaluations. Reproducibility challenges are well-documented (Pineau et al., 2021; Sommerschield et al., 2023); our contribution is the first contamination audit for ancient-language NMT.

3 Methodology

3.1 Data Sources

We use the publicly released data from the hiero-transformer repository:[2]

Table 1 summarizes the filtering pipeline. The TLA's digitization is incomplete: approximately two-thirds of German entries have a translated target but no digitized hieroglyphic source in Gardiner notation—they contain only transliterations, which cannot be used as model input. After filtering for entries with *both* non-empty hieroglyphic source and German target, only **18,669 usable pairs** remain—this is the actual training set for all fine-tuning. Similarly, each test hieroglyphic sentence appears twice (once with a German target, once with English); entries whose German target is empty have translations only in English. This yields just **50 valid test samples**—a size reflecting the fundamental scarcity of digitized hieroglyphic

[2] https://github.com/mattiadc/hiero-transformer

Statistic	Count
Training samples (total)	61,330
German-target (non-empty target)	55,397
German-target (non-empty src & tgt)	18,669
Unique German training targets	15,842
Test samples (total)	150
ea→de, valid	50
Validation samples (total)	125
ea→de, valid	75

Table 1: Dataset statistics. "Valid" = non-empty hieroglyphic source *and* non-empty German target.

data, where expansion requires new expert annotation of archaeological sources.

3.2 Contamination Detection

Normalized exact match. We compare each test target against all 15,842 unique German training targets after normalization (Unicode NFC → lowercase → punctuation removal → whitespace collapse; released as `normalize_translations.py`). Normalization is necessary because Egyptological annotations use bracket conventions (e.g., "[Werde]" vs. "Werde") that mask near-duplicates. This yields the contamination set $\mathcal{C} = \{(s,t) \in \mathcal{D}_{\text{test}} : \text{norm}(t) \in \{\text{norm}(t') : t' \in \mathcal{T}_{\text{train}}\}\}$, partitioning tests into **contaminated** ($|\mathcal{C}| = 16$) and **clean** (34) subsets.

Character n-gram overlap. To capture soft leakage beyond exact matches, we compute the fraction of character 8-grams in each test target appearing in any training target, reporting the number of samples flagged at overlap thresholds from 50% to 100%.

3.3 Models Evaluated

We evaluate five model configurations spanning two training regimes:

Original-regime models. **Released Model**: The publicly available HuggingFace checkpoint of De Cao et al. (2024), an M2M-100 (418M) model fine-tuned on TLA data with Adam (lr=3e-5, fixed schedule, batch size 16). This represents what independent researchers can access.

Script Reproduction: We retrained M2M-100 with its default hyperparameters (epochs=20, batch_size=16, lr=3e-5, Adam, no warmup, no label smoothing). This represents the closest possible replication of the original training procedure.

Our retrained models. The following three models use a modernized training recipe: AdamW optimizer, cosine learning-rate schedule with warmup, label smoothing (ϵ=0.15), weight decay (0.1), dropout tuning (0.2/0.05/0.05 for feedforward/attention/activation), gradient clipping (max norm 1.0), and 5× data upsampling—all applied identically across the three configurations. The effective batch size is 288 (vs. 16 in the original). These choices follow recent best practices for low-resource MT.

M2M-100 Hybrid: M2M-100 (418M) with lr=3e-5 and 1000-step warmup.

M2M-100 Conservative: Identical to Hybrid except lr=1e-5.

mBART-50: mBART-50 (611M) with lr=3e-5 and 500-step warmup, testing whether contamination effects generalize across architectures.

Using a stronger training recipe is deliberate: if contamination inflation persists even with improved optimization, the effect is clearly dataset-inherent rather than an artifact of under-training.

3.4 Evaluation Protocol

We evaluate on three subsets:

- **All**: Complete test set (50 samples)
- **Contaminated**: Samples with targets in training (16 samples)
- **Clean**: Samples with unseen targets (34 samples)

For each subset, we report:

- **BLEU**: Corpus-level BLEU using SacreBLEU (Post, 2018) with case-insensitive scoring (signature: `nrefs:1|case:lc|eff:no|tok:13a|smooth:exp|version:2.6.0`)
- **chrF++**: Character n-gram F-score with word bigrams (Popović, 2015)

We report BLEU with `case:lc` (lowercased internally by SacreBLEU before tokenization) on unmodified model outputs. An ablation comparing case-sensitive versus case-insensitive evaluation confirms that contamination gaps are consistent regardless of text preprocessing.

Generation uses beam search with beam size 10 and maximum length 128, following typical configurations for M2M-100.

4 Results

4.1 Contamination Statistics

Exact matching finds 15/50 (30.0%) verbatim targets; after normalization (lowercasing, whitespace and punctuation standardization), **16/50 (32.0%)** match, where the 16th sample differs from its training counterpart only in parenthetical formatting. The validation set shows a comparable rate (23/75, 30.7%). At the 70% threshold with character 8-grams, 25/50 (50%) of test samples are flagged, confirming that 32% exact match is a conservative lower bound. The consistent rates across test and validation sets indicate a systematic characteristic of the splitting procedure.

4.2 Impact on Automatic Metrics

Table 2 presents translation quality stratified by contamination status.

The gap is substantial across all five models (+29 to +47 BLEU), holding across architectures (M2M-100 vs. mBART-50) and training strategies—the problem is inherent to the dataset. Despite wide 95% CIs on the small subsets (clean: ±7–14 BLEU; contaminated: ±15–20 BLEU), the contaminated and clean intervals do not overlap for any model, confirming the gap is statistically robust. The script reproduction outperforms the released model (+5.2 BLEU) despite identical hyperparameters; the gap is predominantly on contaminated samples (+17.2) versus clean (+2.9), consistent with different checkpoint selection or training seed rather than a systematic quality difference.

COMET-22 Validation. Table 3 confirms the gap using COMET-22 (Rei et al., 2022): contaminated samples score 0.871–0.924 while clean samples score 0.622–0.676 (a gap of 0.23–0.26 points). Model rankings on clean data are identical between BLEU and COMET-22, confirming genuine quality differences beyond contamination effects.

Retrieval Baseline. A retrieval baseline that, given access to the test target, copies the best-matching training target (by character 8-gram coverage) achieves 81.8 BLEU on exact-match items—rivaling the best neural model (77.9)—while scoring 0.0 on clean items. A realistic variant that uses only source similarity (KNN over source embeddings) achieves only 22.0 BLEU on exact-match items. The 59.8-point gap directly measures the contribution of target memorization over source-driven translation.

Granular Contamination Analysis. When we further split the 34 non-exact samples into soft leakage (70–99% 8-gram overlap; n=9) and fully clean (<70%; n=25), BLEU decreases monotonically across all five models—from exact to soft to clean—confirming that performance tracks contamination severity.

Source-Side Overlap. Exact-target samples have higher mean source overlap (0.883) than clean samples (0.543), but this reflects formulaic corpus structure, not causal attribution: **10 of 16 exact-target test items have at least one contaminating training sample with source similarity below 5%**, meaning the matched target was learned from a completely different hieroglyphic inscription. One test item ("Werde fein zerrieben.") has zero source overlap with any training sample yet achieves 100% target overlap. This directly demonstrates target memorization independent of source input.

English Direction. As a sanity check, ea→en evaluation (only 1/50 contaminated) yields 11.5–18.8 BLEU—below German clean scores, confirming contamination-driven inflation in the German direction.

Document-Level Decontamination. We remove every training sentence from the 33 test-source documents (28.4% of training), continue fine-tuning from M2M-100 Conservative for 3,000 steps at lr=3×10^{-6}, and evaluate the best checkpoint (step 300, Val BLEU 37.9). Table 4 reports results: exact-match BLEU drops only 4.6 points (77.9→73.3), while clean BLEU *increases* slightly (39.6→40.3). The residual contamination occurs because **8 of 16 exactly matched test targets persist in training from other source documents**—e.g., "Werde gekocht." (87 training occurrences) appears across many independent papyri. Document-level splitting is thus necessary but not sufficient: target-level deduplication is required.

4.3 Qualitative Analysis

Representative examples illustrate the contrast. Contaminated samples are reproduced verbatim; clean samples exhibit genuine translation errors: "hemsut" (a goddess) → "ausreiß" (escape), "uräen" (sacred cobras) → "rw.t-schlange" (a different serpent type), and "hunderte opfern ihm" (hundreds sacrifice to him) → "er zählt zu 100" (he counts to 100).

Manual categorization of errors in the 34 clean

Model	Subset	n	BLEU	95% CI	chrF++
Released	All	50	37.0	[24.5, 48.8]	56.2
	Contaminated	16	60.3	[37.9, 88.0]	81.5
	Clean	34	30.9	[16.2, 43.6]	48.1
Script Reproduction	All	50	42.2	[28.4, 55.0]	61.2
	Contaminated	16	77.5	[51.5, 93.4]	85.1
	Clean	34	33.8	[20.1, 48.2]	53.7
M2M-100 Hybrid	All	50	46.8	[33.9, 59.2]	63.3
	Contaminated	16	**83.8**	[62.5, 100.0]	**91.9**
	Clean	34	37.3	[23.6, 49.8]	54.0
M2M-100 Conservative	All	50	**47.3**	[35.0, 57.4]	**63.2**
	Contaminated	16	77.9	[59.2, 92.8]	90.6
	Clean	34	**39.2**	[25.0, 50.9]	**54.2**
mBART-50	All	50	41.9	[29.6, 51.8]	59.6
	Contaminated	16	72.8	[55.7, 92.4]	87.3
	Clean	34	33.7	[18.3, 47.1]	50.6

Table 2: Translation quality by contamination status across five models (`case:lc` BLEU, 95% CIs from 1,000 bootstrap resamples). All models show substantial contamination gaps (29–47 BLEU points). Bold indicates best performance per subset.

Model	All	Contam.	Clean
Released (HF)	0.702	0.871	0.622
Exact `train.py`	0.721	0.871	0.650
M2M-100 Hybrid	0.744	**0.924**	0.660
M2M-100 Conservative	**0.753**	0.915	**0.676**
mBART-50	0.732	0.905	0.651

Table 3: COMET-22 scores (`Unbabel/wmt22-comet-da`) on all 50 test samples, the 16 contaminated samples, and the 34 clean samples. Contaminated samples consistently score 0.17–0.26 points higher than clean samples, confirming the contamination-driven inflation observed in BLEU.

predictions reveals: domain-term substitution (35%), entity/deity confusion (24%), syntactic role reversal (18%), number/quantifier errors (15%), and missing content (9%). These error patterns are particularly concerning for DH applications requiring precise entity identification.

4.4 Phrase-Type Analysis

Of the 16 contaminated samples, 13 (81%) are medical formulae—short imperative instructions such as "Werde getrunken" (be drunk) and "Werde fein zermahlen" (be finely ground)—which repeat dozens of times in training. The remaining three are narrative markers (2) and a religious phrase (1). This concentration in formulaic genres confirms that contamination arises systematically from the data structure. The comparable contamination rates between test (32%) and validation (31%) sets further confirm this is a systematic characteristic of the split, not an isolated anomaly.

4.5 Illustrative Mixture Analysis

To illustrate how contamination inflates corpus-level BLEU, we simulate scores at controlled rates using the M2M-100 Conservative model's actual predictions (200 bootstrap trials per rate). At α=0 (fully clean) BLEU=40.2; at our observed rate (α=0.32) BLEU=46.5; at α=0.50 BLEU=52.0; at α=1.0 BLEU=77.9. Under these assumptions, the reported 61.5 BLEU would require a substantially higher contamination rate than the 32% we observe in the released test set, though we cannot estimate a precise rate without access to the original model's predictions and test composition.

5 Discussion

5.1 Sources of Contamination

Three factors produce 32% contamination: (1) **formulaic repetition** in ancient texts—medical instructions, offering formulae, and divine epithets recur across thousands of inscriptions; (2) **source-only deduplication**, which misses cases where different hieroglyphic sequences share the same German target; and (3) **sentence-level splitting**—all 33 unique test documents also appear in training (the highest-overlap document contributes 804 training and 5 test samples), confirming sentence-level rather than document-level partitioning.

5.2 Implications

Our contamination gaps (29–47 BLEU) exceed those in prior work: Kocyigit et al. (2025) found 20–30 point inflation in LLM-based MT, and Magar

Model	All (n=50)	Exact (n=16)	Soft (n=9)	Clean (n=25)
M2M-100 Conservative	47.3	77.9	35.3	39.6
Doc-Clean retrain	45.2	73.3	30.7	**40.3**
Δ	-2.1	-4.6	-4.6	$+0.7$

Table 4: BLEU scores before and after document-level decontamination (M2M-100 Conservative re-trained on 13,365 samples with test-source documents excluded). Exact: 16 samples with 100% 8-gram coverage; Soft: 9 samples with 70–99% coverage; Clean: 25 samples below threshold. Despite removing 28.4% of training data, exact-match BLEU drops only 4.6 points, because 8/16 contaminated targets persist in training from other source documents.

and Schwartz (2022) documented substantial inflation across NLP benchmarks. The gap between reported (61.5) and clean (30.9–39.2) BLEU represents plausible inflation of 22–31 points. Latin, Greek, Akkadian (Gutherz et al., 2023), and other ancient language corpora share the property of formulaic repetition; their benchmarks should be examined similarly.

5.3 Practical Guidance

Our corrected baselines (30.9–39.2 BLEU) indicate that NMT captures overall meaning but contains notable errors. For *corpus triage*, BLEU around 30–39 provides sufficient gist; for *philological analysis*, domain-term substitution (35% of errors) and entity confusion (24%) risk silent misinterpretation. NMT should be deployed as a screening tool with mandatory expert verification.

We recommend that **benchmark creators** check for train-test target overlap and report contamination statistics; **model developers** report performance separately on contaminated and clean subsets; and **consumers** treat reported scores as upper bounds pending contamination analysis.

5.4 Decontaminated Test Set and Splitting Protocol

We release a decontaminated test set of 34 samples (no exact target overlap) spanning medical/magical (41%), religious/funerary (32%), and administrative/literary (27%) content. Clean BLEU (30.9–39.2) represents a *lower bound* on contamination-removal effect and an *upper bound* on true generalization, since residual phrase-level overlap may provide partial memorization benefit.

To prevent contamination, we propose a **document-level splitting protocol**: (1) partition at the level of source texts rather than sentences; (2) deduplicate targets via frequency caps or cluster-and-drop; (3) stratify by genre; and (4) verify zero exact-match target overlap post-hoc. Future work should construct larger test sets from text collections not used in training.

6 Limitations

Test set size. Our clean set (34 samples) yields wide bootstrap CIs ($\pm$7–14 BLEU), though the consistent gap direction across all models confirms robustness. **Soft leakage.** Our primary detection uses exact matching (32%); n-gram analysis shows 50% at the 70% 8-gram threshold, but subset sizes preclude separate evaluation on graduated tiers. **Unavailable original artifacts.** We cannot access the original test set, checkpoint "ea9all", or "test_data" folder, so we cannot directly replicate the reported 61.5 BLEU evaluation. **Length confound.** Contaminated targets are shorter (mean 4.1 vs. 5.8 words), which may partly amplify BLEU inflation—though the retrieval baseline (81.8 BLEU with no translation) confirms the effect is real. **No human evaluation.** We rely on automatic metrics; Egyptologist judgements would strengthen assessment.

7 Conclusion

We reveal that 32% of test samples in a hieroglyphic-to-German NMT benchmark have targets matching training data (50% under 8-gram overlap at 70% threshold). This contamination inflates BLEU by 29–47 points across five model configurations: contaminated samples achieve up to 83.8 BLEU / 0.924 COMET-22 while clean samples achieve only 30.9–39.2 BLEU / 0.622–0.676 COMET-22. Document-level decontamination reduces exact-match BLEU by only 4.6 points, because 8/16 targets persist across other source documents—target-level deduplication is required.

We release a decontaminated 34-sample test set and establish corrected baselines for future work. More broadly, formulaic repetition in ancient corpora makes contamination particularly likely, and

we recommend future benchmarks report contamination statistics and partition results by contamination status.

Reproducibility Statement

All data used in this study is publicly available from the hiero-transformer repository. Our contamination detection code, the decontaminated test set (no exact target overlap, 34 samples), evaluation scripts, a per-item catalog mapping all 50 test items to document IDs, contamination status, training frequency, and source overlap scores, and all retrained model checkpoints are available at `https://github.com/ammarlhassan/hiero-contamination-study`.

Ethics Statement

Our work involves analysis of publicly released models and data. We present our findings as a constructive contribution to reproducibility rather than criticism of prior work—the contamination we identify likely arose from standard preprocessing procedures rather than intentional data manipulation.

Ancient Egyptian texts constitute irreplaceable cultural heritage. We encourage use of our corrected baselines to provide realistic expectations for NMT-assisted Egyptological research, avoiding overreliance on automated systems for scholarly work.

References

Chaymaa Abbas, Nour Shamaa, and Mariette Awad. 2026. Obscuring data contamination through translation: Evidence from arabic corpora. *arXiv preprint arXiv:2601.14994*.

Andrea Barucci, Costanza Cucci, Franco Franci, Marco Loschiavo, and Fabrizio Argenti. 2021. A deep learning approach to ancient Egyptian hieroglyphs classification. *IEEE Access*, 9:123438–123447.

Danlu Chen, Freda Shi, Aditi Agarwal, Jacobo Myerston, and Taylor Berg-Kirkpatrick. 2024. LogogramNLP: Comparing visual and textual representations of ancient logographic writing systems for NLP. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14148–14167. Association for Computational Linguistics.

Mattia De Cao, Nicola De Cao, Angelo Colonna, and Alessandro Lenci. 2024. Deep learning meets egyptology: A hieroglyphic transformer for translating ancient egyptian. *Digital Scholarship in the Humanities*.

Jasper Dekoninck, Mark Niklas Müller, and Martin Vechev. 2024. ConStat: Performance-based contamination detection in large language models. *arXiv preprint arXiv:2405.16281*.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. *arXiv preprint arXiv:2402.15938*.

Maxim Enis and Mark Hopkins. 2024. From LLM to NMT: Advancing low-resource machine translation with claude. *arXiv preprint arXiv:2404.13813*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Morris Franken and Jan C van Gemert. 2013. Automatic Egyptian hieroglyph recognition by retrieving images as texts. In *Proceedings of the 21st ACM International Conference on Multimedia*, pages 765–768.

Gai Gutherz, Shai Gordin, Luis Sáenz, Omer Levy, and Jonathan Berant. 2023. Translating Akkadian to English with neural machine translation. *PNAS Nexus*, 2(5):pgad096.

Muhammed Yusuf Kocyigit, Eleftheria Briakou, Daniel Deutsch, Jiaming Luo, Colin Cherry, and Markus Freitag. 2025. Overestimation in LLM evaluation: A controlled large-scale study on data contamination's impact on machine translation. *arXiv preprint arXiv:2501.18771*.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 157–165.

Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori B. Hashimoto. 2023. Proving test set contamination in black box language models. *arXiv preprint arXiv:2310.17623*.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. 2021. Improving reproducibility in machine learning research: A report from the NeurIPS 2019 reproducibility program. *Journal of Machine Learning Research*, 22(164):1–20.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020. Transforming machine translation: A deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11(1):4381.

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

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585. Association for Computational Linguistics.

Thea Sommerschield, Yannis Assael, John Pavlopoulos, Vanessa Stefanak, Andrew Senior, Chris Dyer, John Bodel, Jonathan Prag, Ion Androutsopoulos, and Nando de Freitas. 2023. Machine learning for ancient languages: A survey. *Computational Linguistics*, 49(3):703–747.

David Tan, Pinzhen Chen, Josef van Genabith, and Koel Dutta Chowdhury. 2026. When flores bloomz wrong: Cross-direction contamination in machine translation evaluation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026)*.

Philipp Wiesenbach and Stefan Riezler. 2019. Multi-task modeling of phonographic languages: Translating middle Egyptian hieroglyphs. In *Proceedings of the 16th International Conference on Spoken Language Translation*.