Deep Past Challenge: decoding 4,000-year-old Akkadian with AI

The Deep Past Challenge is a $50,000 Kaggle competition [Deep Past Initiative +2] **tasking participants with building machine translation models that convert transliterated Old Assyrian Akkadian cuneiform into English** [News Channel Nebraska] [News Channel Nebraska] — representing what organizers call "the ultimate computational frontier" [Deep Past Initiative] in NLP. [Deep Past Initiative] [deeppast] Launched December 16, 2025, with an entry deadline of March 23, 2026, [News Channel Nebraska +2] the competition targets thousands of untranslated clay tablets from ancient Kanesh [Deep Past Initiative] dating to the early second millennium BCE. [Deep Past Initiative] The evaluation metric is the **geometric mean of BLEU and chrF++** [Medium] (Score = √(BLEU × chrF++)), which forces competitors to balance word-level precision with character-level fidelity. [Medium] [Medium] Based on public notebooks and academic benchmarks, **ByT5 (byte-level T5) has emerged as the dominant model architecture**, with MarianMT fine-tuned from Arabic→English as a strong alternative exploiting Semitic language family transfer.

The task: translating Bronze Age merchants into modern English

The competition is hosted by the Deep Past Initiative (DPI), a [News Channel Nebraska] New Haven-based non-profit [News Channel Nebraska +3] co-founded by Dr. Gojko Barjamovic, Dr. Agnete Lassen, [News Channel Nebraska] and Dr. Ruchir Agarwal, [News Channel Nebraska] [EIN Presswire] with financial backing from algorithmic trading firm **XTX Markets**. [News Channel Nebraska +2] The source texts are business records — contracts, letters, loans, receipts, [Deep Past Initiative] courtroom testimonies, and family correspondence — left by Assyrian merchants who built a vast trade network across the ancient Middle East. [Deep Past Initiative] [News Channel Nebraska] Over **22,000 cuneiform tablets** [Deep Past Initiative] exist from this archive, with roughly half untranslated [deeppast] and **fewer than 20** [Deep Past Initiative] **living scholars** [Deep Past Initiative] able to read them. [Deep Past Initiative]

The input is not raw cuneiform images but **transliterated text** — Latin-alphabet representations of cuneiform signs using special conventions including diacritical marks (š, ṣ, ṭ), subscript numbers, and embedded Sumerograms (e.g., "Lugal" for "king"). [aclanthology] [acl-bg] This transliteration format adds complexity because conventions vary across digitization eras, and the cuneiform script itself is described as "more complex than Chinese Hanzi and more ambiguous than Egyptian hieroglyphs," [News Channel Nebraska] [News Channel Nebraska] having been used to write [News Channel Nebraska] over a dozen languages across 3,500+ years. [Deep Past Initiative] [News Channel Nebraska]

The competition uses a **code competition format** (evidenced by the separation of training and inference notebooks across the community), meaning participants likely submit inference notebooks rather than CSV files directly. The competition's subtitle — "Bringing Bronze Age Voices Back to Life" [X +5] — captures both the scholarly and technical ambition.

Evaluation metric balances precision and character similarity

The scoring formula, **Score = √(BLEU × chrF++)**, [Medium] combines two complementary translation metrics via their geometric mean. [Medium] [Medium] This design choice has significant strategic implications for competitors. Medium

BLEU (Bilingual Evaluation Understudy) measures word n-gram overlap Medium — counting matches of 1-word, 2-word, and 3-word sequences between the candidate and reference translations. Medium It is harsh: synonyms score zero (e.g., "dispatch" vs. "send" receives no credit), Medium and missing function words like "the" reduce the score noticeably. Medium BLEU includes a brevity penalty for translations shorter than the reference.

chrF++ measures character-level n-gram similarity with word-boundary awareness. Medium It is more forgiving of morphological variations (plurals, tense), spelling differences, and near-misses. Medium Medium For example, a typo like "sliver" instead of "silver" would be heavily penalized by BLEU but partially rescued by chrF++ since the two words share most characters. Medium

The geometric mean creates a **critical balancing constraint**: if either metric is weak, the overall score drops dramatically even if the other is strong. Medium Medium This means a model that produces creative but lexically imprecise translations (low BLEU, high chrF++) will score poorly, as will one that matches individual words perfectly but scrambles character-level patterns. Competitors must optimize for both simultaneously — favoring models that produce stable, consistent phrasing with high lexical accuracy.

Data landscape and external corpus strategy

While the specific Kaggle data page could not be rendered (Kaggle's JavaScript protection blocks automated access), the competition dataset follows standard Kaggle translation competition structure, likely consisting of **train.csv** (parallel Akkadian-English pairs), **test.csv** (Akkadian text only), and **sample_submission.csv** (format template with id and predicted translation columns).

The training data consists of transliterated Old Assyrian texts paired with expert English translations. Several characteristics make this dataset uniquely challenging:

- **Extremely low-resource**: Even the largest available Akkadian-English parallel corpora contain only tens of thousands of sentence pairs — orders of magnitude smaller than typical NMT training sets
- **Special character handling**: Transliterations contain diacritics Hugging Face (š, ṣ, ṭ), subscript numerals, and non-standard Latin characters that break conventional tokenizers
- **High formulaic repetition**: Business records contain many near-identical legal and commercial formulae, which can inflate apparent model performance aclanthology
- **Sumerograms**: Akkadian texts embed Sumerian logograms that require special treatment aclanthology
- **Fragmentary texts**: Many tablets are physically damaged, yielding incomplete source sequences

Competitors are actively supplementing the competition data with external corpora, strongly suggesting **external data is permitted**. Two community-created Kaggle datasets are prominent: the **"Old Assyrian Extended Corpus"** by user leiwong and the **"Michel Old Assyrian Letters Corpus"** by user manwithacat. Beyond these, critical external resources include:

- **ORACC** (Open Richly Annotated Cuneiform Corpus) — the primary academic source, PubMed Central with sub-corpora including RINAP (Royal Inscriptions of the Neo-Assyrian Period), GitHub SAAo (State Archives of Assyria Online), and RIAo/RIBo (Royal Inscriptions Online) scitepress
- **CDLI** (Cuneiform Digital Library Initiative) — 300,000+ text entries, though most lack translations PubMed Central
- **AICC** (AI Cuneiform Corpus) by Frank Krueger — 130,000 AI-translated texts from CDLI and ORACC, praeclarum model available on HuggingFace at praeclarum/cuneiform
- **Akkademia** project (Gutherz et al.) — GitHub repository with NLP tools and training data for Akkadian GitHub Oxford Academic
- **veezbo/akkadian_english_corpus** — cleaned Akkadian-English parallel corpus on HuggingFace GitHub
- **FactGrid Wikibase** — largest database of Sumerian and Akkadian lexemes mapped to English UC Berkeley School of Informa…

Public notebooks reveal ByT5 as the community favorite

At least **15 public notebooks** have been identified on the competition's code page, spanning EDA, training, and inference. User **leiwong** is the most prolific contributor with at least four notebooks including baseline models and exploratory analyses. Here is the full inventory of identified notebooks:

Notebook	Author	Focus
🏆 DPS - Baseline + Extended Dataset	leiwong	Baseline model with augmented data
Deep Past Challenge - Comprehensive EDA	leiwong	Dataset exploration and statistics
⭐ Deep Past Challenge - EDA + Extended Dataset	leiwong	EDA with external corpus integration
Deep Past Challenge: Starter Notebook	nihilisticneuralnet	Getting-started boilerplate
Deep Past Challenge | Inference	imaadmahmood	Inference pipeline
DPC Starter Train	takamichitoda	Training starter code
Deep Past Challenge | byt5-base | Training	xbar19	ByT5-base fine-tuning
Deep Past Challenge | byt5-base | Inference	xbar19	ByT5-base prediction
Deep Past Challenge | byt5-base | Training v2	sayedathar11	Improved ByT5 training
DeepPast-Akkadian → English-MarianMT	amritanshukush	MarianMT approach
T5_Akkadian_Translation_Model	likithagedipudi	T5-based translation
byt5-akkadian-combined v1.0.6	manwithacat	Combined ByT5 approach
Akkadian T5 Best Inference	manwithacat	T5 inference pipeline
DeepPast-ByT5 SentenceAlign Baseline	djamilabenchikh	ByT5 with sentence alignment

ByT5 (Byte-level T5) dominates the notebook ecosystem with at least 5-6 notebooks. This makes strong technical sense: ByT5 operates on raw UTF-8 bytes rather than subword tokens, completely bypassing the tokenization problems that plague standard models when encountering Akkadian's diacritics and special characters. With **~580M parameters** for ByT5-base and a maximum sequence length of 1,024 bytes, it offers a practical balance between capability and compute requirements.

The separation of training and inference notebooks follows the standard Kaggle code competition pattern, where GPU-intensive training is done offline and lightweight inference runs within submission constraints.

Model strategy: what architectures work for ancient Akkadian

Academic literature and competition notebooks together paint a clear picture of the model landscape. The 2025 paper by Jones & Mitkov (RANLP 2025 Workshop) provides the most rigorous comparative evaluation, testing six fine-tuned models on **95,629 Akkadian-English parallel samples** from ORACC and CDLI: aclanthology

Model	Parameters	BLEU	BERTScore	Inference Speed
Mistral 7B	7B	0.478	0.930	3.57 s/sentence
MarianMT (Arabic→English)	~75M	0.453	0.931	0.22 s/sentence
Krueger T5	250M	0.416	0.930	0.48 s/sentence
Qwen 0.5B	500M	0.403	0.929	0.70 s/sentence
T5-base	250M	0.376	0.914	0.48 s/sentence
MarianMT (Spanish→English)	~75M	0.122	0.842	0.22 s/sentence

The foundational 2023 work by Gutherz et al. (PNAS Nexus) achieved **BLEU4 scores of 36.52 (cuneiform→English) and 37.47 (transliteration→English)** Language Log ResearchGate using a Transformer-based architecture Oxford Academic trained on 50,544 ORACC sentences. Language Log

Several strategic insights emerge from these benchmarks:

Transfer learning from Semitic languages is transformative. MarianMT pre-trained on Arabic→English (MarianAr) acl-bg achieves **BLEU 0.453** — dramatically outperforming the same architecture pre-trained on Spanish→English (MarianEs, BLEU 0.122). aclanthology Arabic's grammatical structure as a fellow Semitic language transfers remarkably well to Akkadian despite millennia of divergence. aclanthology This is the single most impactful architectural choice a competitor can make with MarianMT. acl-bg

Large language models show promise but cost inefficiency. Mistral 7B achieves the highest BLEU acl-bg (0.478) but requires **16× the inference time** of MarianMT aclanthology and demands QLoRA for consumer hardware training. On short sentences (<4 words), decoder-only models actually excel ACL Anthology — Mistral reaches 0.602 BLEU on short inputs. aclanthology

ByT5 addresses the tokenization bottleneck. While not benchmarked in the Jones & Mitkov paper, ByT5's byte-level processing avoids the critical weakness of T5's default tokenizer, which cannot represent all cuneiform transliteration characters (replacing ā, ḫ, ī, ř, š, ṣ, ū with unadorned letters, Hugging Face as noted by Krueger). For this competition's metric — which includes character-level chrF++ — **preserving these characters faithfully matters**.

Bidirectional training as regularization. Krueger's approach of training English→Akkadian alongside Akkadian→English helped stabilize convergence praeclarum and prevent divergence during the 30-epoch training process on 210,247 examples. praeclarum

Preprocessing, tokenization, and data pitfalls to navigate

Successful preprocessing for Akkadian NMT requires addressing several domain-specific challenges. **Normalization** is critical: transliterations from different digitization eras use inconsistent conventions (ASCII approximations vs. proper Unicode diacritics), and standardizing these before training prevents the model from learning spurious distinctions. [praeclarum] [scitepress] Regular expressions should handle artifacts, and all text should be lowercased consistently.

For **tokenization strategy**, the choice is consequential. Standard BPE tokenizers (used by T5, MarianMT) struggle with Akkadian's special characters and may split them unpredictably. Three approaches are viable:

- **Byte-level processing** (ByT5): Operates directly on UTF-8 bytes, completely sidestepping tokenization issues — the most robust option
- **Custom BPE**: Training a SentencePiece or BPE tokenizer on the Akkadian corpus itself, [GitHub] as explored by UC Berkeley's CuneiTranslate project [UC Berkeley School of Informa…]
- **Character-level substitution**: Replacing diacritical characters with ASCII equivalents before tokenization (Krueger's approach), at the cost of information loss [praeclarum]

Key **data pitfalls** include the high formulaic repetition in business texts (which can cause train-test leakage if not carefully managed), [aclanthology] the presence of Sumerograms that require special handling, fragmentary tablet damage creating incomplete sequences, and **domain mismatch** when supplementing Old Assyrian data with Neo-Assyrian or Babylonian corpora from ORACC (different dialects, time periods, and conventions can introduce noise rather than signal). The UC Berkeley CuneiTranslate project specifically found that mixing languages and periods introduced inconsistencies and recommended **segmenting corpora by language and time period**. [UC Berkeley School of Informa…]

Rules, leaderboard, and competition logistics

The competition rules page could not be directly accessed due to Kaggle's JavaScript rendering. However, several rules can be confidently inferred from the competition ecosystem. **External data is almost certainly permitted**, given that multiple high-profile notebooks explicitly reference extended datasets and external corpora without any apparent restriction. **Pre-trained models are clearly allowed**, as notebooks use MarianMT, T5, ByT5, Mistral, and Qwen openly. The competition appears to follow a **code competition format** requiring notebook submissions.

Standard Kaggle competition rules typically include: [The Learning Agency] **5 daily submission limit**, team merger deadlines, restrictions on sharing solutions during the competition, and eligibility exclusions for residents of sanctioned countries (Cuba, Iran, Syria, North Korea, Russia, Sudan, and certain Ukrainian regions). [The Learning Agency] Prize distribution details suggest **up to $15,000 for the top solution** based on social media posts. [X] [X]

The **leaderboard** could not be accessed, and no specific scores were found in any public source — the competition is actively ongoing with approximately 5.5 weeks remaining. No winning solution write-ups exist yet, and competitors are unlikely to share detailed strategies until after the deadline. The **Team and Submissions pages** on Kaggle are user-specific (showing only the authenticated user's own team and submission history) and do not contain public data.

Several **discussion threads** were identified (thread IDs: 663210, 663233, 663357, 663388, 663839, 664079, 664518) but their content could not be rendered. Based on the external Medium article about scoring and the extended dataset notebooks, key discussion topics likely include the scoring metric explanation, external data policies, and approaches to data augmentation.

Conclusion: a blueprint for a competitive solution

This competition sits at a fascinating intersection of cutting-edge NLP and ancient humanities, and several clear strategic principles emerge from the evidence.

Architecture choice matters enormously. ByT5-base offers the best balance of tokenization robustness (critical for chrF++ scoring) and practical compute requirements, while MarianMT fine-tuned from Arabic→English provides the strongest transfer learning signal among encoder-decoder models. aclanthology acl-bg An **ensemble of ByT5 and MarianAr** could optimally balance both BLEU and chrF++ components of the geometric mean metric.

Data augmentation is likely the highest-leverage strategy. The competition's low-resource nature means that carefully curated external data from ORACC (especially Old Assyrian sub-corpora), the Michel Old Assyrian Letters corpus, and the AICC can substantially expand training signal — but domain matching to Old Assyrian specifically (rather than Neo-Assyrian or Babylonian) is critical to avoid introducing noise.

The geometric mean metric demands dual optimization. Models that excel on word-level precision (high BLEU) but introduce character-level errors will be penalized just as heavily as those with good character similarity but poor lexical accuracy. Medium Medium This makes ByT5's character-preserving architecture particularly well-suited, and suggests that **post-processing to ensure consistent, standard phrasing** (avoiding creative synonym choices) will yield disproportionate gains.

The prior academic SOTA of BLEU4 ~37.5 (Gutherz et al. 2023) PubMed Central +2 and the more recent benchmark of BLEU ~0.478 (Mistral 7B, Jones & Mitkov 2025) aclanthology provide reference points, though the competition's Old Assyrian focus, specific test set composition, and combined geometric mean metric make direct comparison imprecise. With 5.5 weeks remaining and an active community building on ByT5, MarianMT, and T5 foundations, the winning solution will likely combine careful external data curation, byte-level or custom tokenization, Semitic language transfer learning, and ensemble strategies that explicitly optimize the BLEU-chrF++ geometric mean.