RespondeoQA: a Benchmark for Bilingual Latin-English Question Answering

Marisa Hudspeth[1] **Patrick J. Burns**[2] **Brendan O'Connor**[1]
[1]Manning College of Information & Computer Sciences, University of Massachusetts Amherst
[2]Institute for the Study of the Ancient World, New York University
{mhudspeth,brenocon}@cs.umass.edu pjb311@nyu.edu

Abstract

We introduce a benchmark dataset for question answering and translation in bilingual Latin and English settings, containing about 7,800 question–answer pairs. The questions are drawn from Latin pedagogical sources, including exams, quizbowl-style trivia, and textbooks ranging from the 1800s to the present. After automated extraction, cleaning, and manual review, the dataset covers a diverse range of question types: knowledge- and skill-based, multihop reasoning, constrained translation, and mixed language pairs. To our knowledge, this is the first QA benchmark centered on Latin. As a case study, we evaluate three large language models–LLaMa 3, Qwen QwQ, and OpenAI's o3-mini–finding that all perform worse on skill-oriented questions. Although the reasoning models perform better on scansion and literary-device tasks, they offer limited improvement overall. QwQ performs slightly better on questions asked in Latin, but LLaMa3 and o3-mini are more task dependent. This dataset provides a new resource for assessing model capabilities in a specialized linguistic and cultural domain, and the creation process can be easily adapted for other languages. The dataset is available at: https://github.com/slanglab/RespondeoQA

Keywords: historical languages, evaluation, question answering

1. Introduction

In recent years, large language models (LLMs) have shown impressive abilities across a wide range of natural language understanding and generation tasks. Yet their performance on many languages, including historical ones like Latin, remains underexplored. Latin occupies a unique position compared to other languages: it is no longer spoken, but has a rich written tradition spanning over two millennia and remains a cornerstone of classical education (Leonhardt, 2013). Because of this history, Latin is a highly frequent language in large-scale archival sources—for example, it is the 5th most prevalent language in two recently released corpora, Common Corpus (Langlais et al., 2025) and Institutional Books (Cargnelutti et al., 2025).

Despite the abundance of Latin textual data, few resources exist for evaluating generative LLMs' capabilities for Latin cultural and language skills. Most Latin-specific datasets are designed for token or sentence level classification tasks more suitable for encoder models (NER, WSD, POS tagging, others), although recent work has begun exploring the abilities of generative LLMs for Latin (Gorovaia et al., 2024; Volk et al., 2024; Marmonier et al., 2025). For machine translation specifically, there are few existing sentence-aligned datasets that are large enough for training or robust evaluation (Martínez Garcia and García Tejedor, 2020; Fischer et al., 2022; Rosenthal, 2023), so automatic sentence alignment methods are an open area of research. Often, automatic sentence alignment for Latin is performed from a digital humanities or corpus analysis perspective, focusing on existing, canonical sources that have been extensively translated and studied (Yousef et al., 2022; Craig et al., 2023). These sentences and their translations are likely to have appeared frequently in LLM pretraining data, raising concerns that LLMs may reproduce memorized translations. Thus, these sentence-aligned datasets may be better suited for training rather than evaluation of LLMs.

Data Source	Year(s)	Type
Exercises in Latin Prosody and Versification	1823	OCR book scan
Latin Grammar and Junior Scholarship Papers	1832	OCR book scan
Certamen	1996–2009	Digital text (MS Word)
National Latin Exam (NLE)	2015, 2020, 2025	Digital text (PDF)

Table 1: Sources of our QA data with year of publication and format type. For Certamen and NLE, we only list the years from which we obtained questions, but both have published materials in other years.

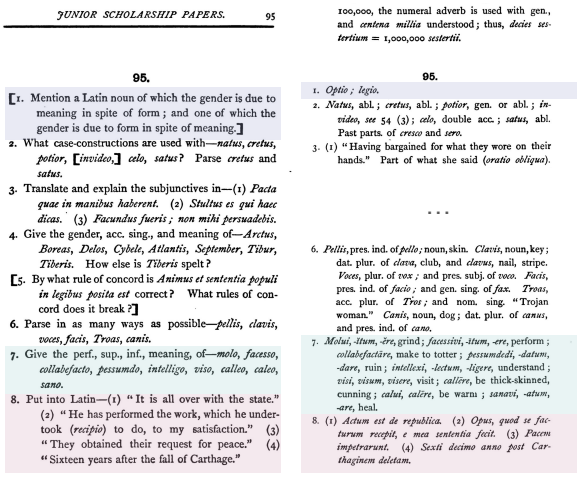
While many multilingual QA benchmarks have been introduced, coverage can be limited for low resource languages, and non-existent for historical languages; as far as we know, none of them include Latin.[1] In addition, these benchmarks generally

[1]The following all exclude Latin: Artetxe et al. (2020); Clark et al. (2020); Lewis et al. (2020); Longpre et al. (2021); Wang et al. (2024); Bandarkar et al. (2024);

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1031–1043
11–16 May 2026.

Content	Source	Question	Answer
(1) Geography	Certamen	Which of these was farthest west in the Roman empire: A: Tarraconensis B: Cappadocia C: Calabria D: Pannonia	A
(2) History	Certamen	Which of the three women, Cornelia, Pompeia, or Calpurnia, was present during the Bona Dea festival that Clodius Pulcher infiltrated while dressed as a woman?	Pompeia
(3) Literature	Certamen	Give the Latin title of the shortest of Plautus' surviving plays.	Curculiō
(4) Mythology	NLE 2015	Quis sum? Ego dē Olympō ad terram dēscendō. Sum nūntius deōrum. Ālās in pedibus meīs habeō. *(Who am I? I descend from Olympus to earth. I am the messenger of the gods. I have wings on my feet.)* A: Neptūnus B: Mercurius C: Iānus D: Mars	B
(5) Vocabulary	*jun-schol*	What is the feminine equivalent of *gener*? *What is the feminine equivalent of *son-in-law*?*	nurus *(daughter-in-law)*
(6) Grammar	Certamen	Dīc fōrmam plūrālem nominis "sceleris." *(Say the plural form of the noun "sceleris.")*	scelerum
(7) Lit. Devices	NLE 2025	What figure of speech is seen in this line from Ennius? Spernitur ōrātor bonus, horridus mīles amātur. A: anaphora B: litotes C: polysyndeton D: chiasmus	D
(8) Read. Comp.	NLE 2020	The tone of Juno's speech throughout this passage is: A: conciliatory B: humble C: persuasive D: condemning	D
(9) Scansion	Certamen	Respondē Latīnē: Dum legis Aeneidem, vidēs haec verba Vergilī: "Conticuēre omnēs intentīque ōra tenēbant." Quot dactylī sunt in versū? *(Respond in Latin: While you read the* Aeneid*, you see this verse by Vergil: "..." How many dactyls are in the verse?)*	duo
(10) Scansion	*lat-pros*	What is the name of each foot in the following line of poetry? Give your answer as a comma-separated list. Intĕ|gēr vī|tæ, scĕlĕ|rīsquĕ | pūrŭs,	Trochee, Spondee, Dactyl, Trochee, Trochee
(11) Scansion	*lat-pros*	Form the following line into hexameter or pentameter verse by changing the position of one word. The word whose position needs to be changed is marked by the * symbol. Ipse dei clypeus terrâ cùm *imâ* tollitur,	Ipse dei clypeus terrâ cùm tollitur imâ.
(12) Translation	*jun-schol*	Put into Latin — "Caius must spare (parco, gerundive) Lucius."	Lucio a Caio parcendum est.
(13) Translation	Certamen	Now translate: sex ursae in silvā erant.	"six bears were in the forest"; "there were six bears in the forest"; "there were six female bears in the forest"; "six female bears were in the forest"

Table 2: Examples of questions across content types, sources, and formats. If the original question or answer is in Latin, we provide an *(italicized translation)* for the reader, which is not in the actual dataset. Details described in §3–5.



JUNIOR SCHOLARSHIP PAPERS. 95

95.

[1. Mention a Latin noun of which the gender is due to meaning in spite of form; and one of which the gender is due to form in spite of meaning.]

2. What case-constructions are used with—*natus, cretus, potior,* [*invideo,*] *celo, satus*? Parse *cretus* and *satus.*

3. Translate and explain the subjunctives in—(1) *Pacta quae in manibus haberent.* (2) *Stultus es qui haec dicas.* (3) *Facundus fueris; non mihi persuadebis.*

4. Give the gender, acc. sing., and meaning of—*Arctus, Boreas, Delos, Cybele, Atlantis, September, Tibur, Tiberis.* How else is *Tiberis* spelt?

[5. By what rule of concord is *Animus et sententia populi in legibus posita est* correct? What rules of concord does it break?]

6. Parse in as many ways as possible—*pellis, clavis, voces, facis, Troas, canis.*

7. Give the perf., sup., inf., meaning, of—*molo, facesso, collabefacto, pessumdo, intelligo, viso, calleo, calco, sano.*

8. Put into Latin—(1) "It is all over with the state." (2) "He has performed the work, which he undertook (*recipio*) to do, to my satisfaction." (3) "They obtained their request for peace." (4) "Sixteen years after the fall of Carthage."

100,000, the numeral adverb is used with gen., and *centena millia* understood; thus, *decies sestertium* = 1,000,000 *sestertii.*

95.

1. *Optio; legio.*

2. *Natus*, abl.; *cretus*, abl.; *potior*, gen. or abl.; *invideo, see* 54 (3); *celo*, double acc.; *satus*, abl. Past parts. of *cresco* and *sero.*

3. (1) "Having bargained for what they wore on their hands." Part of what she said (*oratio obliqua*).

. . .

6. *Pellis*, pres. ind. of *pello*; noun, skin. *Clavis*, noun, key; dat. plur. of *clava*, club, and *clavus*, nail, stripe. *Voces*, plur. of *vox*; and pres. subj. of *voco. Facis*, pres. ind. of *facio*; and gen. sing. of *fax. Troas*, acc. plur. of *Tros*; and nom. sing. "Trojan woman." *Canis*, noun, dog; dat. plur. of *canus*, and pres. ind. of *cano.*

7. *Molui, -itum, -ĕre*, grind; *facessivi, -itum, -ere*, perform; *collabefactare*, make to totter; *pessumdedi, -datum, -dare*, ruin; *intellexi, -lectum, -ligere*, understand; *visi, visum, visere*, visit; *callēre*, be thick-skinned, cunning; *calui, calēre*, be warm; *sanavi, -atum, -are*, heal.

8. (1) *Actum est de republica.* (2) *Opus, quod se facturum recepit, e mea sententia fecit.* (3) *Pacem impetrarunt.* (4) *Sexti decimo anno post Carthaginem deletam.*

Figure 1: A page from *Latin Grammar and Junior Scholarship Papers* (left) and its answer key (right).

do not include bilingual or mixed-language tasks where both the question and answer can interleave two languages. This gap is particularly relevant for Latin, which has long been taught as a second language in bilingual educational settings, and some researchers may need cross-lingual capabilities; for example, to conduct LLM-based analysis of Latin corpora using English-language instructions or questions.

To address these gaps, we introduce a benchmark dataset for question answering and translation in mixed Latin and English settings, comprising approximately 7,800 question–answer pairs. The questions are derived from a diverse set of pedagogical sources, including standardized exams, quizbowl-style trivia, and textbooks spanning the 1800s to the present. These materials represent centuries of pedagogical practice in teaching Latin, ranging from factual recall of mythology or history to more complex reasoning about syntax, translation, and poetic scansion. The resulting dataset is richly annotated with information about the question's format (multiple choice or short answer), content (10 categories), and language; the answer's language; whether the question requires multiple steps of reasoning ("multihop"); and for translation questions,

Xuan et al. (2025); Thellmann et al. (2024).

whether they put constraints on the expected translation.

As a case study, we evaluate three LLMs: two open-source (LLaMa 3 and Qwen QwQ) and one commercial (OpenAI's o3-mini). All exhibit general comprehension but struggle more with skill-oriented questions, such as scansion or translation, which require structured reasoning and linguistic precision. While the reasoning-focused model QwQ shows some advantages on literary and metrical tasks, overall performance is lower than LLaMa's.

Our contributions include:

- **A new mixed-language QA and translation benchmark** for Latin and English, the first of its kind, constructed from education resources spanning multiple centuries.
- **A fine-grained taxonomy of question types** that distinguishes factual knowledge from reasoning skills, enabling more targeted evaluation of LLM capabilities.
- **An evaluation of contemporary open-source LLMs**, revealing gaps in their knowledge and reasoning abilities.

By making this dataset publicly available, we aim to provide a foundation for evaluating and improving LLMs in specialized linguistic and cultural domains.

2. Related Work

Question answering is a staple of Latin learning, though one which recent research suggests the field can benefit from "insight into the realtime comprehension of Latin" (Bextermöller, 2018, pg. 298; see also Kuehnast et al., 2024). Outside the classroom, Latin students have long enjoyed question answering of a different kind, that is "quiz bowl"-style competitions like Certamen that allow students to "demonstrate their knowledge of the ancient peoples, languages, and cultures, and the relationships between those topics and the modern world." (Junior Classical League, 2025). Students can avail themselves of the National Latin Exam for "opportunity to challenge themselves and measure their growth in the study of the Latin language and Greco-Roman culture." (ACL/NJCL, 2024)

Standardized exams are routinely used for benchmarking LLM performance (e.g. SAT, GRE, AP exams in OpenAI, 2023); this is the first foray into LLM-assisted QA applied to these Latin pedagogical assessments. While there is a great deal of NLP work in Latin concerning language modeling and tool development (Riemenschneider and Frank, 2023; Bamman and Burns, 2020), computational work on pedagogical applications is limited (Kuehnast et al., 2024; Schulz, 2021); see also Ross (2023)). More general discussions of ML and AI work involving the Latin language (and other low-resource ancient languages) can be found in Sommerschield et al. (2023).

Bilingual QA is an active area of research, usually with a narrow focus on a particular language pair (Zhang and Wan, 2022; Paschoal et al., 2021; Kalahroodi et al., 2025; Mukanova et al., 2024). Our dataset adds to this growing body of work.

Parallel sentence data for Latin remain limited, which constrains the development and evaluation of neural machine translation (NMT) systems. Several studies focus on automatic sentence alignment methods to construct parallel sentence datasets (Yousef et al., 2022; Craig et al., 2023). These efforts are often situated within digital humanities or corpus linguistics research and typically target well-known literary works with existing translations.

Using either automatically aligned sentences or manually curated alignments, several studies have trained NMT systems for Latin. Prior work includes transformer-based translation systems for Spanish-Latin (Martínez Garcia and García Tejedor, 2020), English-Latin (Rosenthal, 2023), and German-Latin (Fischer et al., 2022). Because Latin translation is a low-resource task, these systems often incorporate techniques such as transfer learning from higher-resource languages (e.g., Italian) or the integration of morphological information to improve performance (Rosenthal, 2023; Fischer et al., 2022). More recent work has also explored the use of generative LLMs for Latin-German translation (Volk et al., 2024).

However, the datasets used in these studies are typically drawn from a narrow set of sources, including the Bible, classical literature, and other widely translated historical texts (Martínez Garcia and García Tejedor, 2020; Rosenthal, 2023). While these corpora are valuable for training NMT systems and for translation studies, their reliance on canonical texts presents challenges for evaluating LLM-based translation systems. Passages from these works and their standard translations are widely circulated in digital corpora, making it likely that they appear frequently in the pretraining data of commercial LLMs. Consequently, evaluations using these datasets may partially reflect memorization rather than a model's ability to generalize to unseen texts.

To partially mitigate this concern, our evaluation dataset is constructed from pedagogical sources, including exams, trivia, and 19th-century textbooks. Although we cannot guarantee that these materials were excluded from LLM pretraining data, they are less widely circulated than canonical literary translations and therefore less likely to appear repeatedly in training corpora. Our dataset also includes features uncommon in existing Latin translation datasets, such as constrained translation exercises

that require the model use specific vocabulary or grammatical constructions, and multiple reference translations.

3. Data Sources

We construct our dataset from four sources, including two textbooks, one set of multiple choice exams, and one set of quizbowl-style trivia questions (Table 1). When looking for potential sources of data, we aimed for a diversity of question types, both in terms of format and content.

Certamen is a quizbowl-style trivia game played competitively by students studying Latin, Greek, and Classical civilizations (Junior Classical League, 2025). Questions cover both language-specific content such as grammar and translation, as well as cultural and historic knowledge. Students can play at three levels of difficulty: novice, intermediate, and advanced.[2]

The National Latin Exam (NLE) is an annual multiple-choice assessment administered to students studying Latin (ACL/NJCL, 2024).[3] As of 2025, there are 8 possible exams offered per year, each around 36-40 questions. The exams include beginner, intermediate, and advanced levels, and dedicated exams for reading comprehension of prose and poetry. Like Certamen, the NLE covers both cultural and linguistic knowledge, dedicating half of the exam to grammar and vocabulary and the other half to knowledge of the Roman world.

Past years' questions from Certamen and NLE are publicly available on their respective websites, but they are copyrighted and not in a structured format ready to be used for NLP applications.

We also sought questions from non-copyrighted historical sources, by searching through scans of Latin textbooks on HathiTrust[4] and the Internet Archive[5] that had corresponding answer keys. We tried to focus on books which had a variety of question types; many books were excluded because they only contained translation questions.

We settled on two 19th century textbooks. The first, *Latin Grammar and Junior Scholarship Papers* (Raven, 1832) contains complex, multipart short answer questions mostly covering vocabulary, grammar, and translation. The second, *Exercises in Latin Prosody and Versification* (Bradley, 1823) has challenging questions related to scansion—describing the formal structure of a poetic line according to its long and short syllables.

Certamen and the NLE are based in the United States, while both textbooks were published in England. As a result, the dataset focuses on English-language and Western pedagogy.

[2] https://www.njcl.org/NJCL-Convention/Convention-Contests/Certamen

[3] https://www.nle.org/

[4] https://www.hathitrust.org/

[5] https://archive.org/

4. Method: Dataset Curation

During each step of our data curation pipeline, if we used a language model for cleanup or annotation, we performed manual review and intervention of its output.

OCR We obtained PDF scans of textbooks and their answer keys from Google Books, and PDFs of the National Latin Exams (NLE) and keys from the NLE website.

For Certamen, we accessed a publicly available Word document containing questions from 1996–2009, which could be exported directly to plain text, eliminating the need for OCR.

We used Gemini-1.5-pro (Gemini Team, 2024) to perform PDF text extraction of the NLE texts and OCR of the textbook scans.

A notable challenge involved the breve (˘), a diacritic used to mark short vowels. Although the breve is not commonly used in Latin writing, it appears extensively in *Exercises in Latin Prosody and Versification*, where the distinction between long and short vowels is essential to the content. OCR inconsistently captured this symbol, resulting in large portions of unusable text.

We manually corrected the accent marks for a small subset of poetry- and scansion-related questions (61 total) from *Exercises in Latin Prosody and Versification*, discarding the rest. Similarly, accent marks were manually corrected in *Latin Grammar and Junior Scholarship Papers* when they were relevant to questions on poetry and scansion.

Alignment of questions to answer keys We used a combination of regular expressions and GPT-4o (OpenAI, 2024) to align questions with their corresponding answer keys.

First, regular expressions were applied to segment each text into a semi-structured format, extracting metadata such as chapters, sections, or grouped numbered text blocks, each representing one or several related questions (for example, multipart questions). The same procedure was applied to both question texts and answer keys, producing roughly aligned pairs based on shared structural information such as section headings.

Certamen materials already contained both questions and answers within the same document, so no separate alignment was needed. In this case, regular expressions were used only to identify sections and extract individual questions.

Exercises in Latin Prosody and Versification presented the greatest challenge, as it interleaved

lessons and explanatory passages with the exercises of interest. We used regular expressions to isolate sections containing exercises and then further subdivided them into individual questions.

For the National Latin Exam (NLE), the formatting was highly consistent, enabling full alignment through regular expressions alone.

For all other sources, we provided Gemini-2.0-Flash with the segmented question and answer text blocks. The model converted these into structured JSON representations, each containing a clearly paired question and answer.

Classify question metadata For each question-answer pair, we used GPT-4o to perform an initial zero-shot classification of several metadata features, then manually reviewed and corrected them.

- **Question format**: multiple choice (MC), or short answer.
- **Question content**: One of 10 possible labels, either knowledge-based (mythology, literature, history, vocabulary, geography) or skill-based (translation, grammar, reading comprehension, scansion, literary devices). This taxonomy is based on lists of topics that Certamen and NLE explicitly intend to cover,[6] and aligns with broader classifications used in Classics education (Canfarotta et al., 2022; Adema, 2019; Vereeck et al., 2024).
- **Question language and answer language**: each could be either English or Latin. Notably, the question language is the language of the instructions or question, not the primary or majority language present in the question text. This distinction is most clear for translation tasks: the question is the instruction ("put into Latin" or "verte in Anglicum") whereas the source language is the language being translated from. We do not explicitly classify the source language for translation questions, but it can be inferred from the answer language (which is equivalent to the target language).
- **Multihop reasoning**: whether the question requires reasoning through intermediate steps in order to reach the final answer. Multihop reasoning is an ongoing area of focus in NLP research, both for training and evaluation (Yang et al., 2018; Tang et al., 2021; Mavi et al., 2024; Zhang et al., 2024). Multihop questions are a common feature of trivia and quizbowl-style datasets in general (Rodriguez et al., 2021; Kabir et al., 2024), so it is unsurprising we found examples in the Certamen dataset. Multihop questions let us test the skill of reasoning LLMs.
- **Constrained translation**: whether the question specifies constraints on the translation, such as requiring a specific lemma or grammatical construction be used.

Some of these attributes—the format, language, and whether a translation is constrained—are straightforward, but the question content and whether a question requires multihop reasoning are more subjective. For example, history and mythology often overlap with literature. While there may be ambiguity in these cases, we are confident in the broader distinction between knowledge and skill-based questions. Similarly, the definition of a multihop question is vague, as it depends on what is considered a distinct step of reasoning.

A subset of the NLE and Certamen questions also have metadata related to question difficulty. We preserve these difficulty labels in the final dataset, but do not perform automatic classification on questions that did not already have a difficulty label. For this reason, we also do not assess LLMs' performance as it relates to question difficulty.

This metadata enables fine-grained analyses of model performance across different aspects of the dataset.

id	question	answer
95.1.2	Is the gender of the Latin noun "legio" determined by form or meaning?	form
95.7.1	Give the perfect form of "molo."	molui
95.7.2	Give the supine form of "molo."	molitum
...		
95.7.29	Give the supine form of "sano."	sanatum
95.8.1	Put into Latin— "It is all over with the state."	Actum est de republica.
...		
95.8.4	Put into Latin— "Sixteen years after the fall of Carthage."	Sexti decimo anno post Carthaginem deletam.

Table 3: Multipart questions from Figure 1 after being broken into standalone questions

Cleanup and Refinement We performed a series of cleanup and refinement steps using GPT-4o and manual review to improve data quality and consistency.

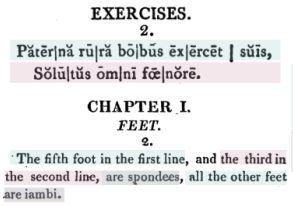
First, **multipart questions** were split into multiple standalone questions, with any references to previous parts disambiguated (Table 3). We then

[6] Certamen and NLE do not have per-question content labels, necessitating zero-shot classification. Both aim to cover language and cultural content. Specifically, Certamen sources questions on grammar and vocabulary, etymology, mottoes, mythology, politico-military history and geography, material culture and social history, and literature (as per the categories listed on the Certamen Source List). The NLE website states it includes questions on "grammar, comprehension, mythology, derivatives, literature, Roman life, history, geography, oral Latin, and Latin in use in the modern world." (URLs accessed March 2026.)

filtered out **unanswerable or invalid questions**, such as those referencing diagrams or missing context, containing multiple equally valid short answers, or containing OCR errors. Since Certamen is meant to be played as a quizbowl-style competition, some questions may instruct the player to perform an action or to comment on actions performed by the moderator. For example, the moderator points at their eye and asks *Quae pars capitis est haec?* (what part of the head is this?). These types of questions were also filtered out.

Next, for ease of evaluation, we filtered out short answer questions whose answers were longer than one whitespace-delimited word.

For short answer translation questions, we filtered out questions whose answers were shorter than three whitespace-delimited words. This filtering does not apply to translation questions in MC format. For these short-answer, long-form translation questions, we create **multiple explicit reference translations** when appropriate. For example, in row 13 of Table 2, our final dataset has 4 gold translations made explicit from the original answer in Certamen: "THERE WERE SIX (FEMALE) BEARS IN THE FOREST / SIX (FEMALE) BEARS WERE IN THE FOREST."



EXERCISES.

2.

Pătēr|nă rū|ră bŏ|bŭs ēx|ĕrcēt | sŭīs,
Sŏlŭ|tŭs ŏm|nī fā|nŏrē.

CHAPTER I.

FEET.

2.

The fifth foot in the first line, and the third in the second line, are spondees, all the other feet are iambi.

id	answer
app.l.2.1	iambus, iambus, iambus, iambus, spondee, iambus
app.l.2.2	iambus, iambus, spondee, iambus

Figure 2: **(left)** Original question and answer from *Exercises in Latin Prosody and Versification*, and **(right)** Answers to the questions, reworded for ease of evaluation.

We also simplified verbose answers for specific question types, particularly prosody and scansion exercises, to make evaluation more reliable (Figure 2).

Some Certamen short-answer questions explicitly gave a candidate list of single-word options, so we found it more natural to reformat them as multiple choice (e.g., "Which of the following Latin nouns does not belong because of gender: cor, agricola, senātus, pēs, leō"). Certamen has no original MC questions, so all 317 MC questions from Certamen were converted from their original SA format using regex.

Finally, we duplicated the scansion-related questions from *Exercises in Latin Prosody and Versification* by translating their English instructions into Latin, ensuring balanced representation across both languages and allowing us to precisely examine the effect of the question language on LLM performance.

5. Dataset Description

Source	MC	1-W SA	Long A.	Total
Certamen	317	4540	970	5827
NLE	855	0	0	855
Lat-Pros	0	0	122	122
Jun-Schol	0	675	350	1025
Total	1172	5215	1442	**7829**

Table 4: Source of data verseus question formats (MC=multiple choice; 1-W SA=one-word short answer; Long A.=long answer).

Our final dataset consists of 7,829 question-answer pairs, with the most common format being 1-word short answer (SA) sourced from Certamen (Table 4). Multiple choice questions have between 3-7 options. Both MC and 1-word SA can be evaluated with accuracy, but long answers include a variety of output types which require separate evaluation strategies (see §6.1).

	Question-Answer Lang				
Content	**En–En**	**La–En**	**En–La**	**La–La**	**Total**
Geography	73	1	107	1	182
History	253	0	685	3	941
Literature	85	2	311	0	398
Mythology	230	1	1283	8	1522
Vocabulary	698	9	733	21	1461
Grammar	192	122	894	106	1314
Lit. Devices	28	1	27	0	56
Read. Comp.	352	7	26	0	385
Scansion	21	20	59	42	142
Translation	854	91	483	0	1428
Total	2786	254	4610	179	**7829**

Table 5: Counts of question–answer language pairs versus question content type. Top rows are **knowledge-based** and bottom rows are **skill-based**.

Table 5 shows the number of QA pairs by language and content type. English questions with Latin answers make up the majority (4610), followed by English questions with English answers (2786). The amount of questions asked in Latin is much smaller, with only 433 total. English questions are also spread nicely across content types, but Latin questions are more sparse.

For translation questions specifically, Table 6 shows there is an over-representation of Latin→English, but the reverse direction still has a sizable amount of examples. For Latin as the target, there are an average of 2.2 reference translations, max of 30; for English, an average of 2.3, max 48. About 13% (166) of the translation questions are constrained, with most of those being

Type	La→En	En→La	Total
Unconstrained	818	285	1103
Constrained	17	146	163
Total	835	431	**1266**

Table 6: Translation direction (src→target) by constraint type for long-form (3+ word) translation questions.

English→Latin (146).

Content	Regular	Multihop	Total
Geography	99	57	156
History	612	280	892
Literature	303	79	382
Mythology	1181	303	1484
Vocabulary	968	266	1234
Grammar	925	77	1002
Lit. Devices	44	2	46
Translation	52	4	56
Scansion	17	2	19
Total	4201	1070	**5271**

Table 7: Counts of regular vs. multi-hop 1-word SA questions and their question content.

Finally, we examine the 1-word short answer questions. About 20% (1070) are multihop (Table 7). The skill-based questions have lower proportions of multihop questions compared to the knowledge-based questions. In particular, geography and history have the highest proportion of multihop questions (37% and 31%, respectively).

We provide examples of questions across all language pairs, content types, formats, and sources in Table 2.

6. Experiments

To illustrate the utility of our dataset to benchmark LLMs, we propose with a set of prompts and evaluation metrics, applied to three current LLMs.

6.1. Experimental Setup

Models We evaluate two open-source LLMs—LLaMa 3.3 (Grattafiori et al., 2024) and Qwen QwQ (Qwen Team et al., 2025; Qwen Team, 2025)—and one commercial model, OpenAI's o3-mini.[7][8] LLaMa3 is a 70 billion parameter instruction tuned model with strong multilingual performance. Qwen QwQ is a smaller, 32 billion parameter model trained from Qwen 2.5 using reinforcement learning (RL) with verifiable outcome-based rewards and a standard reward model. In theory, good reasoning ability could be applicable to our skill-based questions that involve more problem-solving.

Prompts We always provide the system prompt *You are a Classicist with expert knowledge in Greek and Roman history, language, and culture.*

For MC questions, we instruct the model to end their response with the letter of the correct answer. Similarly, for 1-word SA, we ask it to end its response with a single word as its answer.

When prompting the two open models, we use a temperature of 0.6 and top-p of 0.95, the recommended settings for QwQ. These parameters were not tunable for o3-mini using the OpenAI API.

Evaluation Metrics For multiple choice questions, we evaluate the accuracy of the predicted letter choice.

For 1-word short answer questions, we use exact match (EM) accuracy, after normalization (lowercasing, stripping punctuation and whitespace, JV replacement,[9] normalizing macrons, accents, and ligatures). For the majority of questions, normalizing accents will not affect the correctness of the answer. Only questions that ask for vowel quantity to be marked are affected, which is less than 20 questions in the dataset.

For a subset of scansion questions ("feet identification"), we use mean per-item accuracy (see row 10 of Table 2 for an example). Each question gives one line of poetry and asks for the name of each metrical foot in the order it appears in the verse. For a single line, the number of feet typically ranges from 3-6, and partial correctness is allowed.

For another subset of scansion questions ("meter manipulation"), we use accuracy (see row 11 of Table 2). Each question gives a line of poetry and asks that the position of a single word be changed in order to make the verse valid pentameter or hexameter. Since only one word should be moved, we consider it correct (1) if the word is in the correct position and incorrect (0) otherwise.

Finally, for long-form translation questions, we report the BLEU score (Papineni et al., 2002).[10] Although BLEU has received criticism for its low correlation with human judgments (Callison-Burch et al.,

[7] https://openai.com/index/introducing-o3-and-o4-mini/

[8] Version identifiers: `meta-llama/Llama-3.3-70B-Instruct-Turbo`, `Qwen/QwQ-32B`, and `o3-mini-2025-01-31`. Open models were accessed with the Together AI API.

[9] In classical Latin orthography, both I and J are represented by a single letter; same with U and V. It is a common Latin NLP preprocessing step to normalize words to collapse these letter distinctions. For example, the Classical Language Toolkit, an NLP pipeline for pre-modern languages (Johnson et al., 2021), includes a JV replacer and recommends using it during preprocessing.

[10] We use the sacreBLEU implementation (Post, 2018).

2006; Karpinska et al., 2022), the quality of newer, fine-tuned neural metrics for Latin is untested, and there is a lack of gold parallel sentence data available for developing such methods.

6.2. Results

MC and 1-Word SA Accuracy Overall accuracy (excluding long-form translation), is 71.92% for LLaMa3, 68.11% for QwQ, and 68.61% for o3-mini. Performance is much higher on MC formatted questions, with LLaMa at 90.25%, QwQ 90.86%, and o3-mini 91.80%.

All models generally perform better on the knowledge-based questions than the skill-based questions, and LLaMa performs best overall. LLaMa has 74.27% accuracy for knowledge questions, 64.77% for skill; QwQ 71.08% knowledge and 61.63% skill; and o3-mini 69.46% knowledge and 66.76% skill. The gap in performance between knowledge and skill questions is noticeably smaller for o3-mini.

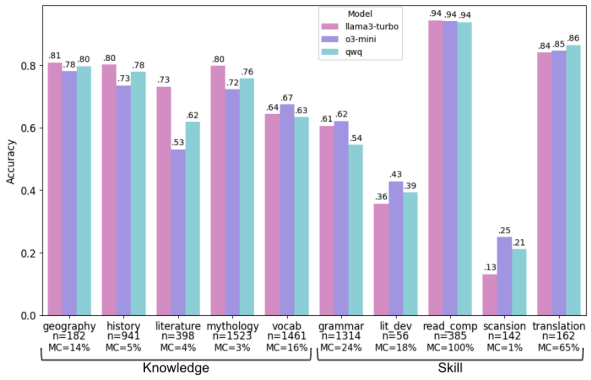


Figure 3: Accuracy by question content. Knowledge categories are on the left and skill categories on the right. Includes both MC and 1-word SA questions.

In Figure 3, LLaMa is the best performer for most knowledge-based content types, and o3-mini is worst performing. The reasoning models, QwQ and o3-mini, show an advantage over LLaMa on literary devices and scansion questions, although the best scansion accuracy (25%, o3-mini) is still far behind the other question types.

High accuracy on the skill-based reading comprehension and translation questions is likely due to their questions being majority MC format.

Effect of Question Language To analyze the effect of the question language on performance, we only report accuracy on Grammar and Scansion questions, since they have enough examples for each language pair.

Although the Latin-English Grammar pairs have the highest accuracy in Figure 4, these are entirely

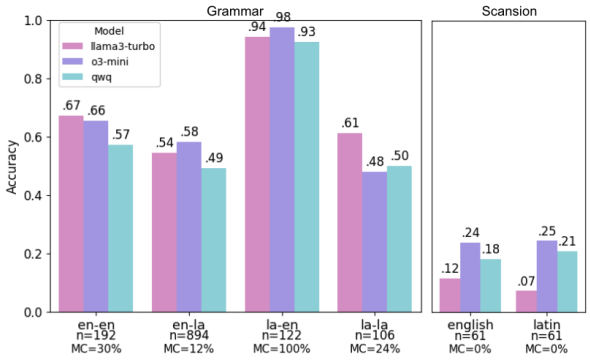


Figure 4: Accuracy by question-answer language pair, for Grammar questions (left), and by question language for Scansion questions (right).

MC formatted questions from NLE.

Keeping the answer language fixed to Latin, LLaMA and QwQ have better performance when a Grammar question is asked in Latin rather than in English. This gap is larger for LLaMa (61% La-La, 54% En-La) than for QwQ (50% La-La, 49% En-La). For o3-mini, this effect is reversed, with a 10% drop in accuracy.

Similarly, QwQ performs better on scansion questions asked in Latin (21%) than those asked in English (18%). However, this is reversed for LLaMa, with 7% accuracy on scansion questions asked in Latin and 12% on those asked in English. Accuracy is about the same for o3-mini for English (24%) and Latin (25%) questions.

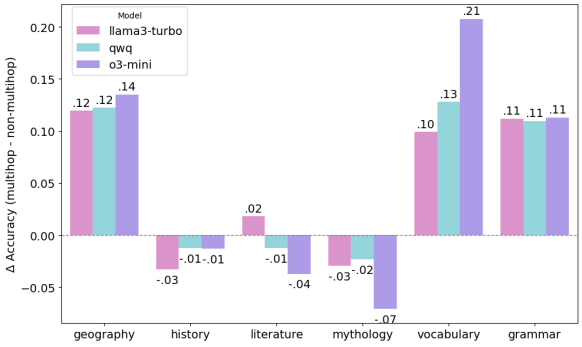


Figure 5: Difference in accuracy: multihop - regular for 1-word SA questions.

Multihop questions Surprisingly, all the LLMs performed better overall on the multihop questions (Figure 5). Some content types (history, literature, mythology) showed mild regressions on multihop questions. These mixed results may be due to the subjectivity of what qualifies as a distinct step of reasoning. In addition, tasks that require multiple steps of reasoning for a human may have answers that are directly stated in LLM training data. For example, a grammar question may ask for a word to be inflected to a particular form. A human may

need to think through multiple steps to come to the correct answer (what declension/conjugation is the word, what are the correct endings for that declension/conjugation, what is the particular ending the question is asking for, and how does that ending combine with this lemma?), but an LLM may have seen all inflected forms neatly formatted together in its training data.

	LLaMA 3		QwQ		o3-mini	
Setting	**Lat**	**Eng**	**Lat**	**Eng**	**Lat**	**Eng**
Unconst.	25.50	45.41	18.53	39.91	23.42	43.29
Const.	20.88	37.46	21.52	17.95	27.25	34.68
Overall	23.71	**45.25**	19.53	39.27	**24.67**	43.14

Table 8: BLEU scores for each translation setting and target language.

Long-form Translation Overall, LLaMa is the best model for translating into English, and o3-mini is best for translating into Latin.

There are very few constrained translation questions where English is the target language, so results are less valid.

LLaMa performs worse on constrained translations across each target language setting. QwQ sees a large drop on constrained translation when English is the target, but we observed this was caused by the model not following directions and leaving explanations in its final answer.

QwQ and o3-mini perform better on the constrained Latin translations than the unconstrained ones, outperforming LLaMa. If a model is correctly able to follow the given constraint, then the space of possible translations is smaller, so it should be easier to provide a translation closer to the reference(s).

7. Discussion and Future Work

Reasoning abilities are beneficial for some skill-based tasks (scansion, literary devices) but are unable to compensate for poorer foundational knowledge. Considering the added computational cost, it is unnecessary to use reasoning models for most tasks we tested. We also observed QwQ's reasoning ability sometimes prevented it from coming to an answer at all, getting stuck in reasoning loops. However, o3-mini, the other reasoning model, did not have the same issue.

In this paper, we use basic prompting strategies, but more targeted techniques may need to be developed to improve performance in certain areas such as vocabulary, grammar, and especially scansion.

All models lag behind when translating into Latin versus English. More work should investigate this gap, as well as the effect of the instruction language and the type of constraint present in the instruction.

Future work to flesh out sparser content types and language pairs in the dataset would be especially valuable. Additionally, questions with long, phrase- or sentence-length answers could be added, and automatic evaluation methods could be tested for these questions.

Although we use this dataset for evaluation only, it could also be used for training of MT systems or instruction tuning of larger generative models.

8. Conclusion

We present the first benchmark for QA and translation in mixed Latin–English settings, built from over 7000 questions spanning two centuries of pedagogical materials and capturing a wide spectrum of linguistic and reasoning challenges. Our evaluation of three large language models reveals that even strong general-purpose models struggle with skill-based and linguistically precise tasks. We hope this resource will support future research on multilingual and historical language understanding, and serve as a blueprint for building comparable resources in other low-resource settings.

9. Ethics

Our dataset is derived from publicly available materials, but some subsets are copyrighted and have distinct terms of use and access.

At the time of writing, we do not plan to redistribute the portions of our dataset sourced from Certamen. The Junior Classical League (JCL) has agreed to host the Certamen portion of our dataset on its website along with their archived Certamen questions.

We will host the subset of our data sourced from NLE. The ACL/JCL National Latin Exam does not allow these materials to be used for generation of profit.

The most up-to-date access to the dataset and details on terms of use will be maintained at: https://github.com/slanglab/RespondeoQA.

10. Limitations

It is possible that our questions exist in LLM pretraining data. However, the performance of the tested models still has room for improvement. Even if our data was seen by the models during training, it is also unlikely to have seen answers aligned to the questions.

Some combinations of question types, content, and languages are sparsely represented in our dataset, so a robust evaluation of performance is

not yet possible. We try to limit evaluations to categories that have enough examples.

11. Acknowledgments

We would like to thank the UMass NLP group for their feedback and commentary on this project. This material is based in part upon work supported by National Science Foundation award 1845576 (CAREER). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

12. Bibliographical References

ACL/NJCL. 2024. About Us. National Latin Exam website.

Suzanne Adema. 2019. Latin learning and instruction as a research field. *Journal of Latin Linguistics*, 18(1-2):35–59.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

David Bamman and Patrick J. Burns. 2020. Latin BERT: A Contextual Language Model for Classical Philology. ArXiv: 2009.10053.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The Belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Delaram Bextermöller. 2018. Reading Latin and the need for empirical research: A psycholinguistic approach to reading comprehension in Latin. *Journal of Latin Linguistics*, 17(2):281–300.

C. Bradley. 1823. *Exercises in Latin Prosody and Versification*. A.J. Valpy.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Daniela Canfarotta, Crispino Tosto, and Raquel Casado-Muñoz. 2022. Development of key competences through latin and greek in secondary school in italy and spain. *Journal of Classics Teaching*, 23(45):13–21.

Matteo Cargnelutti, Catherine Brobston, John Hess, Jack Cushman, Kristi Mukk, Aristana Scourtas, Kyle Courtney, Greg Leppert, Amanda Watson, Martha Whitehead, and Jonathan Zittrain. 2025. Institutional Books 1.0: A 242B token dataset from Harvard Library's collections, refined for accuracy and usability. ArXiv:2506.08300 [cs].

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Caroline Craig, Kartik Goyal, Gregory R. Crane, Farnoosh Shamsian, and David A. Smith. 2023. Testing the Limits of Neural Sentence Alignment Models on Classical Greek and Latin Texts and Translations. In *Workshop on Computational Humanities Research*.

Lukas Fischer, Patricia Scheurer, Raphael Schwitter, and Martin Volk. 2022. Machine translation of 16Th century letters from Latin to German. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 43–50, Marseille, France. European Language Resources Association.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Svetlana Gorovaia, Gleb Schmidt, and Ivan P. Yamshchikov. 2024. Sui generis: Large language models for authorship attribution and verification in Latin. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 398–412, Miami, USA. Association for Computational Linguistics.

Aaron Grattafiori et al. 2024. The Llama 3 herd of models.

Kyle P. Johnson, Patrick J. Burns, John Stewart, Todd Cook, Clément Besnier, and William J. B. Mattingly. 2021. The Classical Language Toolkit: An NLP framework for pre-modern languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the*

11th International Joint Conference on Natural Language Processing: System Demonstrations, pages 20–29, Online. Association for Computational Linguistics.

Junior Classical League. 2025. 2025 NJCL Certamen Rules.

Tasnim Kabir, Yoo Yeon Sung, Saptarashmi Bandyopadhyay, Hao Zou, Abhranil Chandra, and Jordan Lee Boyd-Graber. 2024. You make me feel like a natural question: Training QA systems on transformed trivia questions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20486–20510, Miami, Florida, USA. Association for Computational Linguistics.

Mohammad Javad Ranjbar Kalahroodi, Amirhossein Sheikholselami, Sepehr Karimi, Sepideh Ranjbar Kalahroodi, Heshaam Faili, and Azadeh Shakery. 2025. PersianMedQA: Evaluating Large Language Models on a Persian-English Bilingual Medical Question Answering Benchmark.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. DEMETR: Diagnosing evaluation metrics for translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9540–9561, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Milena Kuehnast, Konstantin Schulz, and Anke Lüdeling. 2024. Development of basic reading skills in Latin. *Cogent education*, 11(1). Publisher: Sprach- und literaturwissenschaftliche Fakultät.

Pierre-Carl Langlais, Carlos Rosas Hinostroza, Mattia Nee, Catherine Arnett, Pavel Chizhov, Eliot Krzystof Jones, Irène Girard, David Mach, Anastasia Stasenko, and Ivan P. Yamshchikov. 2025. Common Corpus: The Largest Collection of Ethical Data for LLM Pre-Training. ArXiv:2506.01732 [cs].

Jürgen Leonhardt. 2013. *Latin: Story of a World Language*. Harvard University Press.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Malik Marmonier, Rachel Bawden, and Benoît Sagot. 2025. Explicit learning and the LLM in machine translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31372–31422, Suzhou, China. Association for Computational Linguistics.

Eva Martínez Garcia and Álvaro García Tejedor. 2020. Latin-Spanish neural machine translation: from the Bible to saint augustine. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 94–99, Marseille, France. European Language Resources Association (ELRA).

Vaibhav Mavi, Anubhav Jangra, and Adam Jatowt. 2024. Multi-hop question answering. *Found. Trends Inf. Retr.*, 17(5):457–586.

Assel Mukanova, Alibek Barlybayev, Aizhan Nazyrova, Lyazzat Kussepova, Bakhyt Matkarimov, and Gulnazym Abdikalyk. 2024. Development of a Geographical Question- Answering System in the Kazakh Language. *IEEE Access*, 12:105460–105469.

OpenAI. 2023. GPT-4 Technical Report. ArXiv:2303.08774 [cs].

OpenAI. 2024. Gpt-4o system card.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

André F. A. Paschoal, Paulo Pirozelli, Valdinei Freire, Karina V. Delgado, Sarajane M. Peres, Marcos M. José, Flávio Nakasato, André S. Oliveira, Anarosa A. F. Brandão, Anna H. R. Costa, and Fabio G. Cozman. 2021. Pirá: A Bilingual Portuguese-English Dataset for Question-Answering about the Ocean. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 4544–4553, New York, NY, USA. Association for Computing Machinery.

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, Belgium, Brussels. Association for Computational Linguistics.

Qwen Team, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

John Hardy Raven. 1832. *Latin Grammar and Junior Scholarship Papers*. Rivingtons.

Frederick Riemenschneider and Anette Frank. 2023. Exploring large language models for classical philology. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15181–15199, Toronto, Canada. Association for Computational Linguistics.

Pedro Rodriguez, Shi Feng, Mohit Iyyer, He He, and Jordan Boyd-Graber. 2021. Quizbowl: The case for incremental question answering.

Gil Rosenthal. 2023. Machina cognoscens: Neural machine translation for latin, a case-marked free-order language. Master's thesis, University of Chicago.

Edward A. S. Ross. 2023. A New Frontier: AI and Ancient Language Pedagogy. *Journal of Classics Teaching*, 24(48):1–19. Publisher: Cambridge University Press.

Konstantin Schulz. 2021. Natural Language Processing for Teaching Ancient Languages. In *Teaching Classics in the Digital Age*, pages 37–48. Universitätsverlag Kiel, Kiel.

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*, pages 1–45.

Yixuan Tang, Hwee Tou Ng, and Anthony Tung. 2021. Do multi-hop question answering systems know how to answer the single-hop sub-questions? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3244–3249, Online. Association for Computational Linguistics.

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, and Mehdi Ali. 2024. Towards multilingual llm evaluation for european languages.

Alexandra Vereeck, Evelien Bracke, Katja De Herdt, and Mark Janse. 2024. Revered and reviled. an outline of the public debate regarding classical language education. *Journal of Classics Teaching*, 25(50):101–115.

Martin Volk, Dominic Philipp Fischer, Lukas Fischer, Patricia Scheurer, and Phillip Benjamin Ströbel. 2024. LLM-based machine translation and summarization for Latin. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 122–128, Torino, Italia. ELRA and ICCL.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, Felix Juefei-Xu, Foutse Khomh, Osamu Yoshie, Qingyu Chen, Douglas Teodoro, Nan Liu, Randy Goebel, Lei Ma, Edison Marrese-Taylor, Shijian Lu, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. 2025. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1513–1532, Suzhou, China. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Tariq Yousef, Chiara Palladino, David J. Wright, and Monica Berti. 2022. Automatic translation align-

ment for Ancient Greek and Latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 101–107, Marseille, France. European Language Resources Association.

Jiahao Zhang, Haiyang Zhang, Dongmei Zhang, Liu Yong, and Shen Huang. 2024. End-to-end beam retrieval for multi-hop question answering. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1718–1731, Mexico City, Mexico. Association for Computational Linguistics.

Yunxiang Zhang and Xiaojun Wan. 2022. BiRdQA: A Bilingual Dataset for Question Answering on Tricky Riddles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11748–11756.