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, UMass Amherst
[2]Institute for the Study of the Ancient World, New York University

Motivation

- Generative LLMs' performance on Latin is underexplored, despite being the 5th most prevalent language in recent large pretraining corpora
- Existing multilingual QA benchmarks exclude Latin entirely
- We introduce **~7,800 QA pairs** from pedagogical sources spanning 1823–2025, with rich metadata across 10 content types

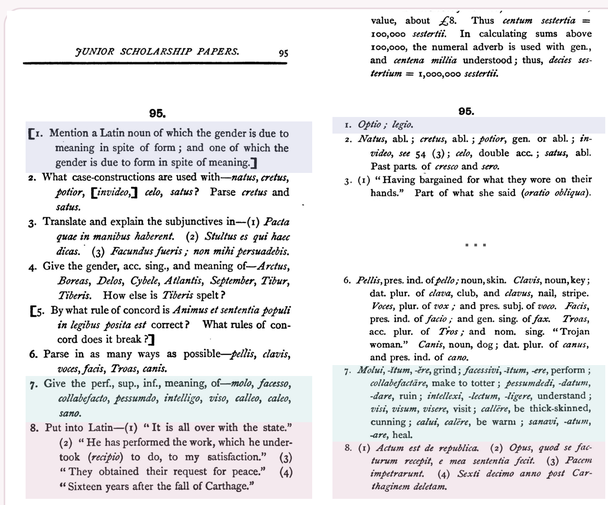
Dataset Sources & Statistics

Data Source	Year(s)	Type	MC	1-W SA	Long Ans	Total
Exercises in Latin Prosody and Versification	1823	OCR scan	0	0	122	122
Latin Grammar and Junior Scholarship Papers	1832	OCR scan	0	675	350	1025
Certamen	1996–2009	Digital (Word)	317	4540	970	5827
National Latin Exam	2015, 2020, 2025	Digital (PDF)	855	0	0	855
			1172	5215	1442	**7829**

Number of Question Types per Source

Dataset Curation

1. **OCR & Extraction** — use Gemini-1.5-pro to perform OCR on PDF scans + digital text
2. **Question–Answer Alignment** — combination of regex + GPT-4o to align questions to their answers
3. **Metadata Classification** — format, content, language, multihop, constrained translations
4. **Multipart Splitting** — Break multipart questions into standalone items



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, collabefacio, 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."

Question Page

value, about £8. Thus *centum sestertia* = 100,000 *sestertii.* In calculating sums above 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*).

* * *

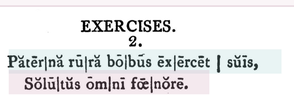
6. *Pellis,* pres. ind. of *pello;* noun, skin. *Clavis,* noun, key; dat. plur. of *clava,* club, and *clavus,* nail, stripe. *Voces,* plur. of *vox;* and pres. subj. of *voco. Facis,* pres. ind. of *facio;* and gen. sing. of *fax. Troas,* acc. plur. of *Tros;* and nom. sing. "Trojan woman." *Canis,*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

Answer Key

id	question	answer
95.1.1	Is the gender of Latin noun "optio" determined by meaning or form?	meaning
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.

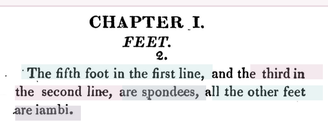
Expanded questions

5. **Answer Refinement** — Simplify verbose answers for easier evaluation



EXERCISES.
2.
Pătēr|nă rū|rā bō|būs ĕx|ercēt | sŭīs,
Sŏlŭ|tŭs ōm|nī fœ|nŏrē.

Question Page



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.

Answer Key

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

Simplified answers

6. **Reference Translations** — create multiple explicit reference translations

"THERE WERE SIX (FEMALE) BEARS IN THE FOREST / SIX (FEMALE) BEARS WERE IN THE FOREST." →

1. "six bears were in the forest"
2. "there were six bears in the forest"
3. "there were six female bears in the forest"
4. "six female bears were in the forest"

7. **Filtering** — Remove unanswerable, diagram-dependent, and ambiguous questions

Question Taxonomy

*10 content types × 4 language pair combinations. Top rows are **knowledge-based;** bottom rows are **skill-based**.*

	Content Type	En→En	La→En	En→La	La→La	Total
KNOWLEDGE	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
SKILL	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**

Question Language→Answer Language

Key Findings

- All models struggle more with skill-based questions than knowledge-based ones
- LLaMA 3 leads overall (71.9%); reasoning models (QwQ, o3-mini) show advantages only on literary devices & scansion
- Best scansion accuracy is only 25% (o3-mini): significant room for improvement
- Question language matters: the reasoning models perform better on scansion questions asked in Latin, whereas LLaMa does better on English questions

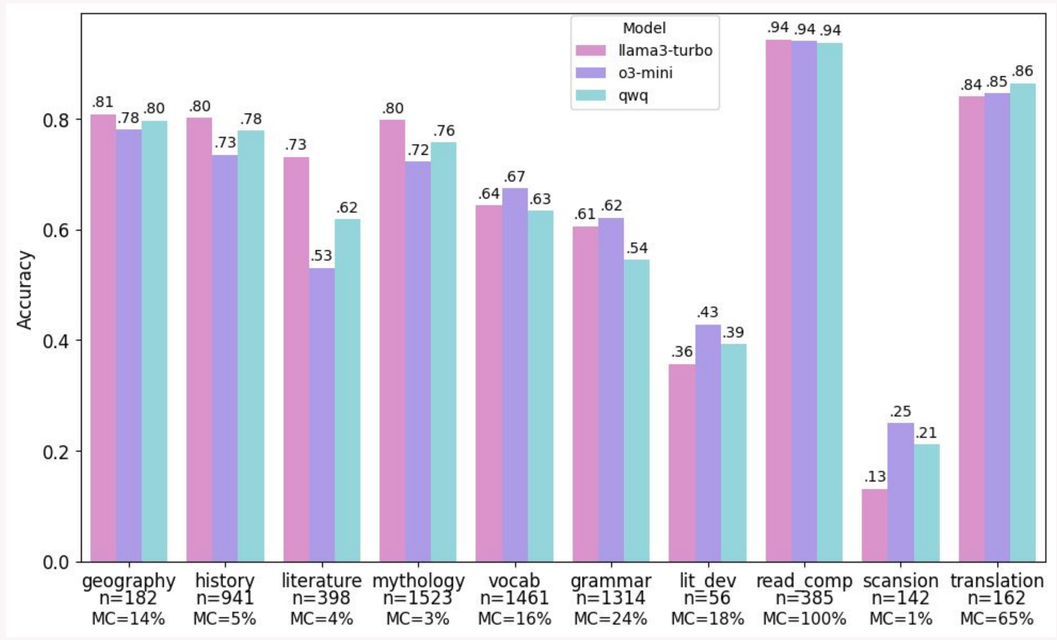
Results: Overall Accuracy

All models: knowledge > skill.

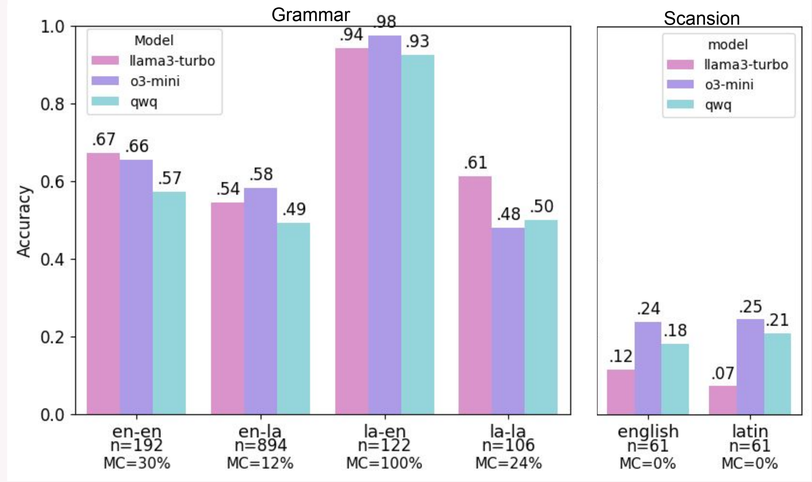
MC accuracy ~90% across all models.

	Overall	Knowledge	Skill
LLaMA 3	**71.92**	**74.27**	64.77
QwQ	68.11	71.08	61.63
o3-mini	68.61	69.46	**66.76**

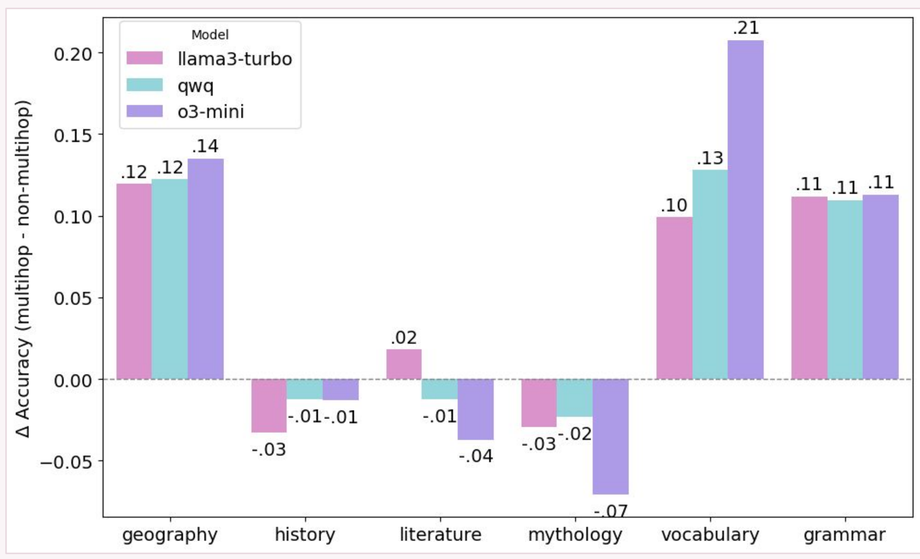
Accuracy by Content Type



Effect of Question Language



Multihop Question Accuracy



Translation Results (BLEU)

	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

LREC 20 Palma 26

University of Massachusetts Amherst

NYU | ISAW