Evaluating Large Language Models for Automatic Generation of EpiDoc TEI from Leiden Transcription

Jun Ogawa（The University of Tokyo）

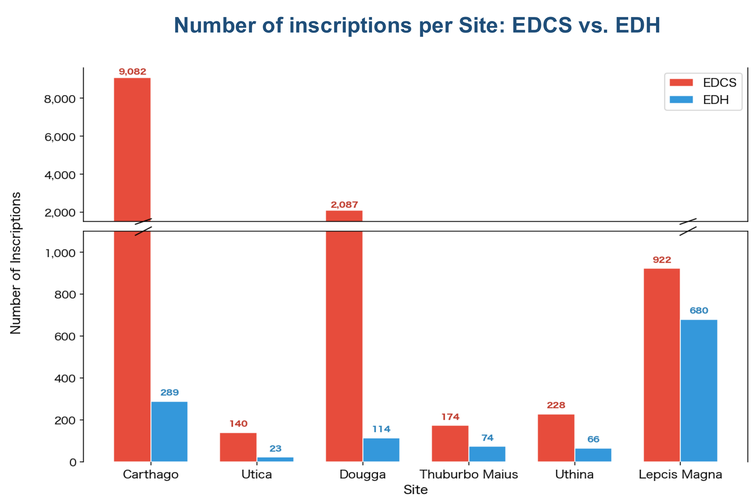
10th Epigraphy.info Workshop@Graz, 24 Mar 2026

東京大学 THE UNIVERSITY OF TOKYO

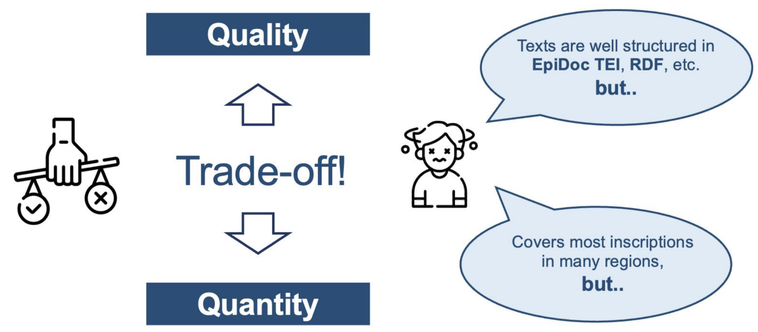
Background

Field trip in Tunisia (Jan. 2025)

- Research project on the state of agriculture in **Roman North Africa**
- Since epigraphic sources are abundant in the region, digitization can dacilitate their efficient search and analysis



For the whole province of Africa Proconsularis, EDCS: **33448**, EDH: **4440**



Motivation

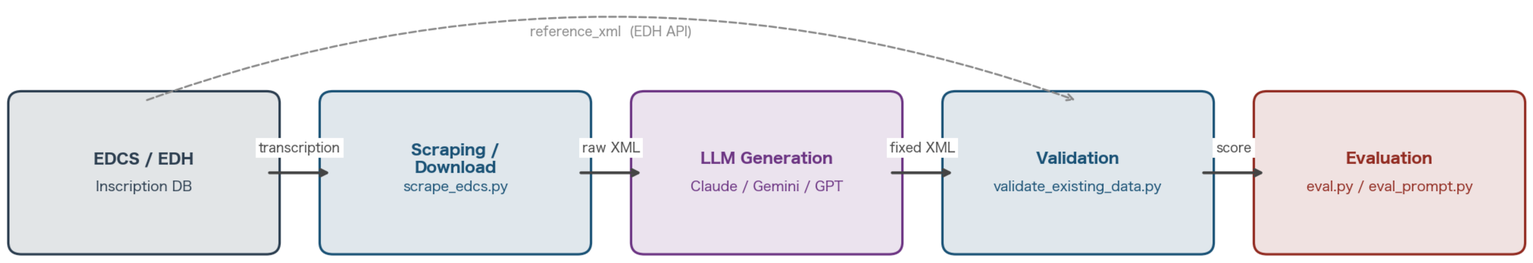
Constructing a large **EpiDoc XML corpora from un-structured transcriptions** to fill the gap between ‘Quality’ and ‘Quantity’

- ✓ While most transcriptions (e.g., EDCS) broadly follow Leiden conventions, **unique rules, irregularities, inconsistency** also exist
  - ➢ unclear character amount → “ A[3] / [3 Ut]hina ”
  - ➢ irregular notational forms → “ duo<v=B>iral(ic) V ”
  - ➢ uncertain letters or restorations → “ Ger[manicus(?) ”
  - ➢ etc..

Rule-based approach [1] has significant limitations ⟷ **LLMs** may better handle such insconsistencies

- ✓ we experimentally sampled **EDH transcriptions coupled with corresponding EpiDoc** editions, to evaluate the usability & suitability of LLMs for automatic generation

Workflow



Generation

	A	B	C	D
1	hd-no.	transcription	work status	province / Italic region
2	HD048091	------] splendi[dissimus? ------	provisional	Africa Proconsularis
3	HD048092	------]A[---] / [--- Ut]hina [------	provisional	Africa Proconsularis
4	HD048093	[--- Domi]tian[i] / [Caesa]ris Au[g	provisional	Africa Proconsularis
5	HD048094	------] Tra[iani] Germ(anici) Da[c(	provisional	Africa Proconsularis
6	HD048095	[--- Tr]aianu[s ---] / [--- t]r(ibunic	provisional	Africa Proconsularis
7	HD048096	[H]a[d]ria[n---]	provisional	Africa Proconsularis
8	HD048097	Di[vi ---] / Hadria[n---]	provisional	Africa Proconsularis
9	HD048098	------ Had]rian[i ---] / [------] // [	provisional	Africa Proconsularis
10	HD048099	------ Se]ptimi S[everi ---] / [--- A	provisional	Africa Proconsularis
11	HD048100	[---]A[---] / [---]P [---]/le[--- c]o	provisional	Africa Proconsularis
12	HD048101	------ Iovi] Optim[o Maximo ------	provisional	Africa Proconsularis
13	HD048102	Gen(io) c[ol(oniae?) ---] / ex tes[t	provisional	Africa Proconsularis
14	HD048103	[---] Au[---] / [------] / A[---] / M	provisional	Africa Proconsularis

69 inscriptions from Uthina, a colony in Africa Proconsularis, downloaded from EDH

- The same transcriptions and prompts were provided to three different models: **GPT 5.2**, **Gemini 3 Flash**, and **Claude 4.6 Sonnet**
- Two types of prompts were used:
  - ✓ Providing an URL of the online EpiDoc Guidelines [2] along with the basic structure of EpiDoc XML markup (Fig.1)
  - ✓ The Guidelines supplemented with a set of markup examples (Fig. 2) extracted from the EpiDoc cheat sheet [3]

The structure of the <text> element, including the <text> tag itself, should be as follows:

```
---------- --
<text>
  <body>
    <div type="edition">
      <ab>
        <lb n="1"/>First line of inscription text
        <lb n="2"/>Second line of inscription text
        ...
      </ab>
    </div>
  </body>
</text>
---------- ---------
Please follow these rules:
・Do not include xml:id.
・Do not start with <?xml version=\"1.0\" ?>; begin with <text>.
・Do not include spaces between tags.
```

Fig.1: Basic EpiDoc/TEI structure

```
Line breaks:
abc / abc → <lb n="1"/>first line <lb n="2"/>second line
1. first line
2. second line → <lb n="1"/>first line <lb n="2"/>second line

Word divided across lines:
abc/def → <lb n="1"/>αβγ <lb n="2" break="no"/>δεζ
αβγ-
δεζ → <lb n="1"/>αβγ <lb n="2" break="no"/>δεζ
1. αβγ
2.- δεζ → <lb n="1"/>αβγ <lb n="2" break="no"/>δεζ

Text divisions:
abc // abc → <div type="textpart" subtype="face" n="r">…</
div><div type="textpart" subtype="face" n="v">…</div>
<D=.r … =D>
<D=.v … =D> → <div type="textpart" subtype="face" n="r">…</
div><div type="textpart" subtype="face" n="v">…</div>

Letters ambiguous outside of their context:
ạḅc̣ → <unclear>αβ</unclear>
αβ̣ → <unclear>αβ</unclear>
```

Fig.2: A part of examples given

Example categories

- ✓ Line breaks, Words divided across lines, Text divisions, ambiguous letters, Erased, Lost characters restored by editors, Lacuna, Line lost, Omitted letters, Abbreviation, Numeral, etc.

Validation

- ✓ Checking tag consistency such as the missing of end tags
- ✓ Removing redundant elements such as unnecessary spaces

Evaluation

F1 value of reproduced XML element path in the output; Normalized Levenshtein similarity of plain text extracted from both XMLs; F1 value of reproduced attributes and their values in the output

Model	Overall Score	Structural Match Rate (w=0.3)	Content Similarity (w=0.5)	Attribute Coverage Rate (w=0.2)
Simple Prompt				
Claude 4.6 Sonnet	86.1%	82.7%	90.9%	85.6%
Gemini 3 Flash	90.1%	85.3%	95.4%	87.5%
GPT 5.2	81.7%	77.1%	86.3%	81.0%
Detailed Prompt				
Claude 4.6 Sonnet	90.4% (+4.3)	88.9% (+6.2)	93.2% (+2.3)	85.9% (+0.3)
Gemini 3 Flash	91.7% (+1.6)	92.2% (+6.9)	94.2% (-1.2)	86% (-1.5)
GPT 5.2	87.8% (+6.1)	86.4% (+9.3)	91.1% (+4.8)	81.8% (+0.8)

Evaluation Results: Comparison between generated XML and EDH XML

➢ Interpretation

Gemini appears most compatible with EpiDoc XML encoding by default, while Claude and GPT benefit more from encoding examples, suggesting a strong influence of prompt design

➢ Comparison among **different models**

HD048095: [--- Tr]aianu[s ---] / [--- t]r(ibunicia) pot(estate) X[---] / [---] ded[it ------

① Reference XML encoding given by EDH

```
<lb n="2"/>
<gap reason="lost" extent="unknown" unit="character"/>
<expan>
  <abbr>
    <supplied reason="lost">t</supplied>
    r
  </abbr>
  <ex>ibunicia</ex>
</expan>
```

② XML generated with **Gemini**

```
<lb n="2"/>
<gap reason="lost" extent="unknown" unit="character"/>
<expan>
  <abbr>
    <supplied reason="lost">t</supplied>
    r
  </abbr>
  <ex>ibunicia</ex>
</expan>
```

➢ Gemini(②) produced XML that was fully identical to the EDH reference(①)

③ XML generated with **Claude**

```
<lb n="2"/>
<gap reason="lost" extent="unknown" unit="character"/>
<expan>
  <abbr>t</abbr>
  <ex>ribunicia</ex>
</expan>
```

➢ Claude(③) failed to extract ‘supplied’ information for ‘[--- t]r(ibunicia

④ XML generated with **GPT**

```
<lb n="2"/>
<gap reason="lost" extent="unknown" unit="character"/>
t]
<expan>
  <abbr>r</abbr>
  <ex>ibunicia</ex>
</expan>
```

➢ ‘]’ remained unconverted as well as the range of abbreviated word is incorrect.

➢ Comparison between **simple & detailed prompts**

HD048094: ------]A[---] / [--- Ut]hina [------

```
<lb n="2"/>
<gap reason="lost" extent="unknown" unit="character"/>
Uthina
<gap reason="lost" extent="unknown" unit="character"/>
</ab>
```

Prompt **only with the Guidelines** (Claude)

```
<lb n="2"/>
<gap reason="lost" extent="unknown" unit="character"/>
<supplied reason="lost">Ut</supplied>
hina
<gap reason="lost" extent="unknown" unit="character"/>
```

Prompt **with encoding examples** (Claude)

➢ According to the transcription, the letters ‘Ut’ seem to be restored by the editor and that is correctly encoded only with the detailed prompt

HD048094: ------] Tra[iani] Germ(anici) Da[c(ici) ---] / [--- ordi]nis Ut[hin---] / [---] patron[o coloniae ------

```
Tra
<supplied reason="lost">iani</supplied>
Germ
<expan>
  <abbr>an</abbr>
  <ex>ici</ex>
</expan>
Da
```

```
Tra
<supplied reason="lost">iani</supplied>
<expan>
  <abbr>Germ</abbr>
  <ex>anici</ex>
</expan>
Da
<supplied reason="lost">
```

➢ Looking at the abbreviation encoding for ‘Germ(anici)’, the prompt including encoding examples successfully corrects errors observed in the simpler prompt, including ‘Germ’ under <expan> element

Discussion

1. Application for EDCS transcriptions

While both EDH and EDCS fundamentally apply Leiden style editing, practical descriptions significantly differ..

EDH052494 ↔ EDCS26500936

[--- Iun]oni [---] / [--- Se?]cundus [---] / [--- pisci?]nas incru[statas? ---] / [---]ium promi[serat ---]

]ON[3] / [3 Iu]cundus [3] / [3 pisci]nas incru[statas(?) 3] / [3]IV promi[serat

To what extent can LLM-based generation absorb descriptive variations for the identical information?

```
<supplied reason="lost">
  <gap reason="lost" quantity="3" unit="character"/>
  pisci
</supplied>
nas incru
<supplied reason="lost" cert="low">statas</supplied>
<gap reason="lost" quantity="3" unit="character"/>
<lb n="4"/>
<gap reason="lost" quantity="3" unit="character"/>
```

```
<gap reason="lost" extent="unknown" unit="character"/>
<lb n="3"/>
<gap reason="lost" extent="unknown" unit="character"/>
<supplied reason="lost" cert="low">pisci</supplied>
nas incru
<supplied reason="lost" cert="low">statas</supplied>
<gap reason="lost" extent="unknown" unit="character"/>
<lb n="4"/>
<gap reason="lost" extent="unknown" unit="character"/>
ium promi
```

Fig.3: EDH – EDCS comparison on ‘detailed prompt with Gemini’ generation

The LLM (Gemini) **correctly interprets the editorial marks,** such as ‘gap’ and ‘supplied’, of the text **regardless of descriptive differences**. + Generation using EDCS transcriptions performs even better in handling the quantity of textual gaps.

2. ‘Ground Truth’ Problem

EDH XML editions contain a certain number of inconsistencies, which may affect the overall evaluation results.. (e.g. **HD048091**)

Creating a set of qualified Ground Truth data is necessary to conduct a more accurate evaluation for LLM-based EpiDoc generation

Conclusion

SUMMARY

- Evaluation pipeline for LLM-base XML generation is designed [4]
- LLMs, especially Gemini 3 Flash can efficiently convert epigraphic texts even with much variations and irregularities
- Encoding examples can enhance the accuracy of the LLM generation, particularly in its structural features

FUTURE TASKS

- Consider using and evaluating Local LLMs for the sustainability
- Construct a more sophisticated pipeline such as Agentic Framwork
- Extend the scope of the encoding to entities and events / relations
- Analyze the outputs to accumulate knowledge about the strength and error patterns of each LLM for future improvement of the pipeline

[1] E.g. *Atlas Patrimonii Caesaris*’s EpiDoc Converter: https://patrimonium.huma-num.fr/atlas/epidoc-converter/ (accessed 14 March 2026). [2] EpiDoc version: post 9.8 dev, List of All Transcription Guidelines: https://epidoc.stoa.org/gl/latest/app-alltrans.html (accessed 14 March 2026). [3] EpiDoc Leiden Quick Reference: https://svn.code.sf.net/p/epidoc/code/trunk/guidelines/msword/cheatsheet.pdf (accessed 14 March 2026). [4] S. Strutz (2026) ‘A Multi-Dimensional Evaluation Framework for Assessing LLM Performance in TEI Encoding,’ *Journal of Open Humanities Data*, 12: 39, pp. 1-20, DOI: https://doi.org/10.5334/johd.484.

Acknowledgement: This work was supported by JSPS KAKENHI Grant Number 24K15674. The author would also like to express sincere gratitude to Dr. Ikki Ohmukai, an associate professor at the University of Tokyo, for his valuable advice in the preparation of this poster.