

IA générative pour l'encodage de grands corpus lexicographiques en XML-TEI

Maritza Beatriz García Rodríguez

Denis Vigier, Ludovic Moncla



Plan

Introduction

Autour de l'encodage XML-TEI

- Choix d'un standard d'encodage XML-TEI
- Schéma d'encodage idéal

Méthodologie

- Niveaux de balisage
- Expérimentations
- Évaluation

Résultats et Observations

Conclusion et perspectives

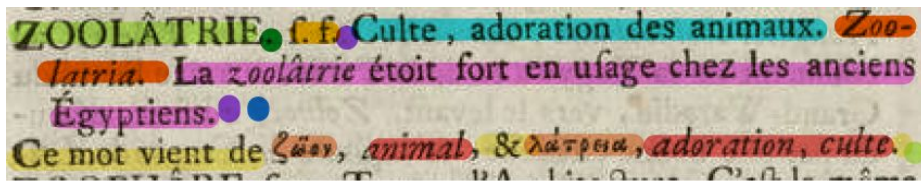
Introduction

- Structure TEI :
 - héritière des projets qui ont précédé ;
 - capable de répondre aux spécificités du Trévoux.
- De l'encodage idéal (réalisé par l'humain) vers l'encodage "au kilomètre" (selon les prédictions de quelques modèles d'IAg)
- Coût : 9.8M tokens (\$8.3)

Choix d'un standard d'encodage XML-TEI

- TEI P5
- TEI P5 chapitre 10 “Dictionaries”
- TEI-Lex 0

Schéma d'encodage idéal



Entrée “Zoolâtrie”, DUFLT 1743. Disponible
sur [DUFLT Philologic](#).

```
<entry xml:id="" type="" xml:lang="">
  <form type="lemma">
    <orth></orth>
    <pron></pron>
  </form>
  <metamark function="lemmaDelimiter"></metamark>
  <gramGrp>
    <gram type="" expand="" norm=""></gram>
  </gramGrp>
  <sense>
    <usg></usg>
    <def></def>
    <cit type="translation">
      <quote></quote>
    </cit>
    <cit type="example">
      <quote></quote><bibl></bibl>
    </cit>
    <ref><bibl></bibl></ref>
    <xr>
      <lbl></lbl>
      <ref></ref>
    </xr>
  </sense>
  <milestone unit="paragraph" rend="leftShift"/>
  <etym>
    <oRef type="etymon"></oRef> <gloss></gloss>
  </etym>
  <note></note>
  <entry xml:id="" type="relatedEntry" xml:lang="">
    [...]
  </entry>
</entry>
```

Méthodologie

- Niveaux de balisage
- Expérimentations
- Évaluation

Niveaux de balisage

1er Niveau

- entrées principales
- sous-entrées
- sens
- formes orthographiques
- informations grammaticales
- retours chariot
- marques d'usage (domaines, temporalité, géographie, style de langue)

```
<entry type="mainEntry">
  <form type="lemmaGrp">
    <form type="lemma">
      <orth>ZOOÏATRIE</orth>
    </form>
    <metamark function="lemmaDelimiter">.</metamark>
    <gramGrp>
      <gram type="pos" expand="substantif"
        norm="NOUN">s.</gram>
      <gram type="gen" expand="feminin"
        norm="Fem">f.</gram>
    </gramGrp>
  </form>
  <sense>Culte, adoration des animaux. Zoolatria.
    La zoolâtrie étoit fort en usage chez
    les anciens Égyptiens.
  <milestone unit="paragraph"
    rendition="leftShift"/>
  Ce mot vient de , animal, & ; , adoration,
  culte.</sense>
</entry>
```

2ème Niveau

- définitions
- traductions
- exemples
- références bibliographiques
- renvois
- références

```
<entry type="mainEntry">
  <form type="lemmaGrp">
    <form type="lemma">
      <orth>ZOOÏATRIE</orth>
    </form>
    <metamark function="lemmaDelimiter">.</metamark>
    <gramGrp>
      <gram type="pos" expand="substantif"
        norm="NOUN">s.</gram>
      <gram type="gen" expand="feminin"
        norm="Fem">f.</gram>
    </gramGrp>
  </form>
  <sense>
    <def>Culte, adoration des animaux.</def>
    <cit type="translation" xml:lang="la">
      <quote>Zoolatria.</quote>
    </cit>
    <cit type="example">
      <quote>La zoolâtrie étoit fort en usage
        chez les anciens Égyptiens.</quote>
    </cit>
    <milestone unit="paragraph"
      rendition="leftShift"/>
    Ce mot vient de , animal, & ; ,
    adoration, culte.</sense>
</entry>
```

3ème Niveau

- étymologies
- étymons
- gloses
- notes

```
<entry type="mainEntry">
  <form type="lemmaGrp">
    <form type="lemma">
      <orth>ZOOÏATRIE</orth>
    </form>
    <metamark function="lemmaDelimiter">.</metamark>
    <gramGrp>
      <gram type="pos" expand="substantif"
        norm="NOUN">s.</gram>
      <gram type="gen" expand="feminin"
        norm="Fem">f.</gram>
    </gramGrp>
  </form>
  <sense>
    <def>Culte, adoration des animaux.</def>
    <cit type="translation" xml:lang="la">
      <quote>Zoolatria.</quote>
    </cit>
    <cit type="example">
      <quote>La zoolâtrie étoit fort en usage
        chez les anciens Égyptiens.</quote>
    </cit>
    <milestone unit="paragraph"
      rendition="leftShift"/>
    <etym><lbl>Ce mot vient de</lbl> ,
    <gloss>animal</gloss> , & ; ,
    <gloss>adoration, culte.</gloss></etym>
  </sense>
</entry>
```

Expérimentations

Traitement du premier niveau d'encodage

- prompt unique
- *prompt chaining*

LLMs (API OpenRouter)

- google/gemma-3-27b-it
- google/gemini-3.1-flash-lite-preview
- openai/gpt-5-mini
- openai/gpt-5.4-mini

Techniques de *prompt engineering*

- *zero-shot* (ZS)
- *few-shot* (FS)
- *few-shot* avec instructions réduites (FS-R)

Évaluation

Protocole

- Validité formelle
- Adéquation structurelle
- Fidélité au texte source
- Robustesse

Analyse

- Qualitative
- Quantitative
 - taux de validité des balises
 - <milestone>
 - <usg type="dom">
 - <usg type="other">
 - précision de segmentation
 - <entry type="mainEntry">
 - <entry type="relatedEntry">

Vérité de terrain

- 3 niveaux d'encodage
- 10 pages du *Trévoux* 1743 :
 - volume 1 : pages 453-454 et 2001-2002 ;
 - volume 2 : pages 37-38 et 1785-1786 ;
 - volume 3 : pages 5-6 et 7-8 ;
 - volume 4 : pages 131-132 et 489-490 ;
 - volume 5 : page 505-506 ;
 - volume 6 : page 1003-1004

Résultats

Moyenne de la f-mesure par type de balise (mainEntry, relatedEntry, milestone, usg_dom, usg_other).

Paramètres :

- modèle : gpt-5-mini
- stratégie : *few-shot* (FS)

traitement	mean_f1 mainEntry	mean_f1 relatedEntry	mean_f1 milestone	mean_f1 usg_dom	mean_f1 usg_other
prompt unique	0,94	0,20	0,46	0,10	0,20
prompt chaining	0,95	0,24	0,57	0,61	0,51

Résultats

Moyenne de la f-mesure par type de balises et par stratégie de *prompting*.

Paramètres :

- modèle : gpt-5-mini
- traitement : *prompt chaining*

stratégie	mean_f1 mainEntry	mean_f1 relatedEntry	mean_f1 milestone	mean_f1 usg_dom	mean_f1 usg_other
FS	0,95	0,24	0,57	0,61	0,51
FS-R	0,94	0,23	0,07	0,15	0,21
ZS	0,95	0,20	0,33	0,15	0,00

Résultats

Moyenne de la f-mesure par type de balises et par modèle.

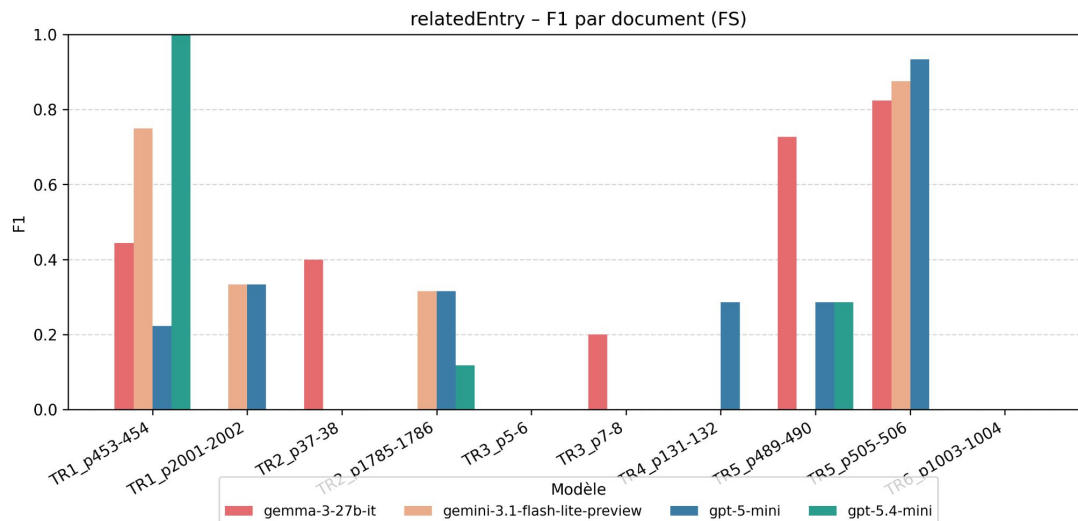
Paramètres :

- traitement : *prompt chaining*
- stratégie : *few-shots*

model	mean_f1 mainEntry	mean_f1 relatedEntry	mean_f1 milestone	mean_f1 usg_dom	mean_f1 usg_other
gemma-3-27b-it	0,76	0,26	0,44	0,00	0,00
gemini-3.1-flash-lite-preview	0,86	0,23	0,67	0,46	0,27
gpt-5-mini	0,95	0,24	0,57	0,61	0,51
gpt-5.4-mini	0,93	0,14	0,17	0,52	0,17

Résultats

Variabilité dans la prédiction des sous-entrées (f-mesure) par fichier et modèle



Observations

Intervention dans le texte du modèle gpt-5-mini avec une configuration *few-shot* et un traitement *prompt-chaining*

Suppression

```
<entry type="mainEntry">
  <form type="lemmaGrp">
    <form type="lemma"><orth>CHARRETIER</orth></form> , ou
    <form type="variant"><orth>CHARTIER</orth></form>
    <form type="inflected"><orth>ière</orth></form>
    <metamark function="lemmaDelimiter">,</metamark>
    <gramGrp>
      <gram type="pos" expand="substantif" norm="NOUN">s.</gram>
    </gramGrp>
  </form>
```

Entrée “Charretier”, DUFLT 1743.

Duplication

```
<entry type="mainEntry">
  <form type="lemmaGrp">
    <form type="lemma"><orth>CHARRETIN</orth></form>
    <form type="variant"><orth>CHARRETEIN</orth></form>
    <metamark function="lemmaDelimiter">,</metamark>
    <gramGrp>
      <gram type="pos" expand="substantif" norm="NOUN">s.</gram>
    </gramGrp>
  </form>
  <sense>ou CHARRETEIN, comme écrit Liger. Espèce de charrette [...]
```

Entrée “Charretin”, DUFLT 1743.

Conclusion et perspectives

- Choix :
 - traitement du premier niveau : *prompt chaining*
 - stratégie de prompt engineering : *few-shot*
 - modèle : gpt-5-mini ?
- Complexifier le traitement du *prompt chaining* en combinant les modèles les plus efficaces selon la tâche.
- Élargir l'encodage vers le deuxième et le troisième niveau de balisage.
- Quelle place pour des plus petits modèles ?

Bibliographie

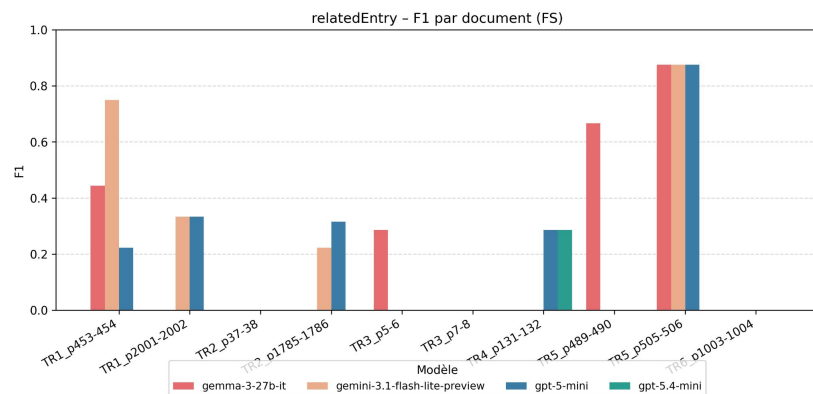
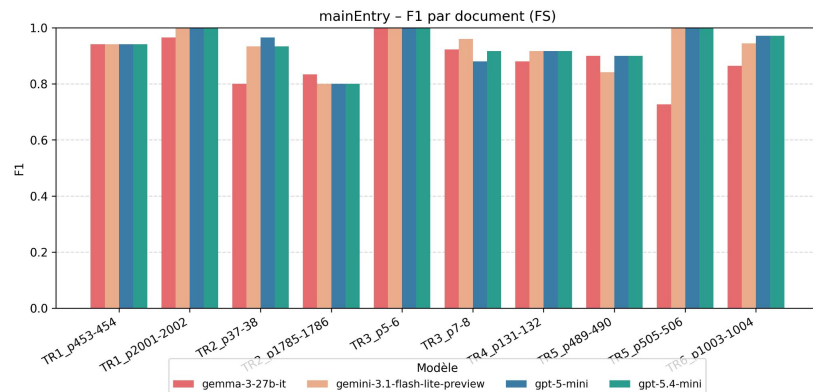
- Bowers, J. et Romary L. (2017) « Deep encoding of etymological information in TEI », *Journal of the Text Encoding Initiative*, 10 <https://inria.hal.science/hal-01296498v2>.
- Bowers, J., Herold, A., Tasovac, T. et Romary L. (2022) « TEI Lex-0 Etym : Toward Terse Recommendations for the Encoding of Etymological Information », *Journal of the Text Encoding Initiative*, Rolling Issue <https://doi.org/10.4000/jtei.4300>.
- Burnard, L. (2020) « What is TEI Conformance, and Why Should You Care ? », *Journal of the Text Encoding Initiative*, Issue 12, Juillet 2019-Mai 2020. En ligne depuis le 6 décembre 2018. URL : <http://journals.openedition.org/jtei/1777> ; DOI: <https://doi.org/10.4000/jtei.1777>.
- Brenon, A. (2024). « Encoding the Specificities of Encyclopedias ». In Javier Martin Arista; Ana Elvira Ojanguren López. *Structuring Lexical Data and Digitising Dictionaries*, Brill, pp.36-62.
- Chambat, A. et Rey, C. (2026). De la rétroconversion à la mise en réseau : dictionnaires et révolution numérique. *Langue française*, 229(1), 91-108. <https://doi-org.bibelec.univ-lyon2.fr/10.3917/lf.229.0091>.
- Galleron, I. (2023) « L'entropie informationnelle dans le *Dictionnaire Universel* (1701). Le cas des étymologies », in Robert Alessi, Marcelo-Vitali-Rosati (dir.), *Les Éditions critiques numériques* (édition augmentée), Les Presses de l'Université de Montréal, Montréal <https://parcoursnumeriques-pum.ca/12-editionscritiques/chapitre8.html>.
- Leca-Tsiomis, M. (2023), *La guerre des dictionnaires. Le Trévoux, aux sources de l'Encyclopédie*, CNRS éditions.
- Le Guern, M. (1983), «Le « Dictionnaire » de Trévoux (1704)», *Cahiers de l'Association internationale des études françaises*, 35, 51-68. DOI : <https://doi.org/10.3406/caief.1983.2402>.
- Salgado, A., Romary, L., Costa, R. *et al.*, (2024), « The Morais dictionary : following best practices in a retro-digitized dictionary project », *International Journal of Humanities and Arts Computing*, 18.1, p. 125-147. DOI : 10.3366/ijhac.2024.0325.
- Turcan, I. (ed.) (2009), *Quand le “Dictionnaire de Trévoux” rayonne sur l'Europe des Lumières*, Paris, L'Harmattan.
- Vigier, D. (2022), « L'esprit encyclopédique moderne en France entre 1690 et 1902. Présentation », *Langue française*, 214(2), 9-26. <https://doi.org/10.3917/lf.214.0009>.
- Wei J. et al. (2022) Chain-of-thought prompting elicits reasoning in large language models.Proceedings of the 36th International Conference on Neural Information Processing Systems NIPS'22 <https://dl.acm.org/doi/10.5555/3600270.3602070>.
- Wionet, C., Tutin, A. (2001), *Pour informatiser le Dictionnaire Universel de Basnage (1702) et Trévoux (1704) ; approche théorique et pratique*, Paris : H. Champion.

Annexes

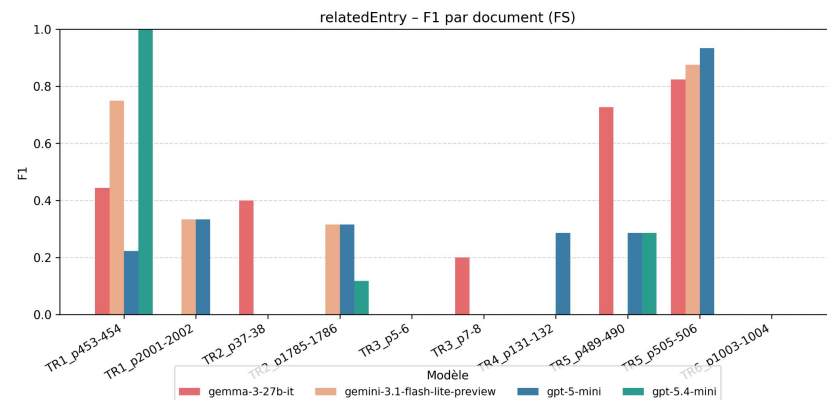
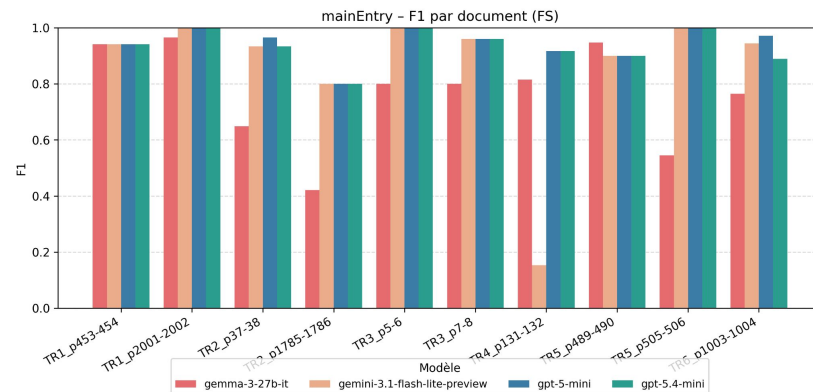
[Voir les balises de plus près](#)

Différence entre la prédiction des entrées principales et les sous-entrées par page et modèle

Prompt unique

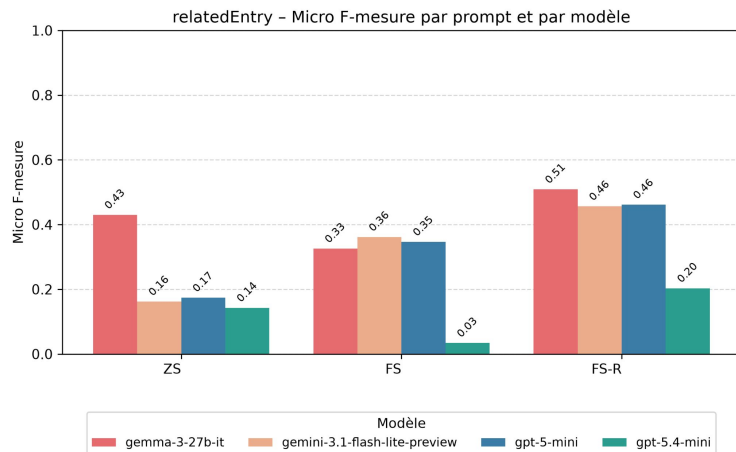
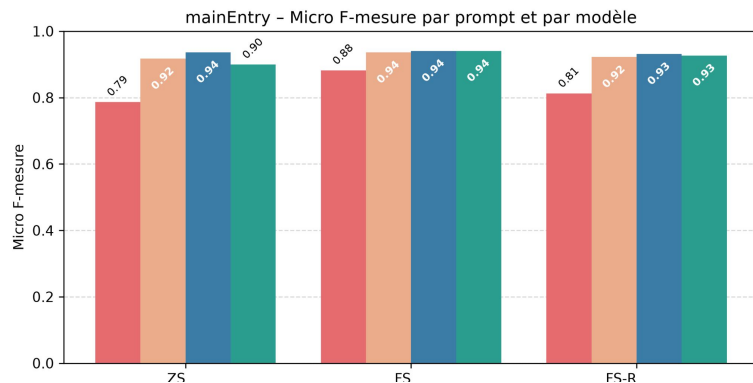


Prompt chaining

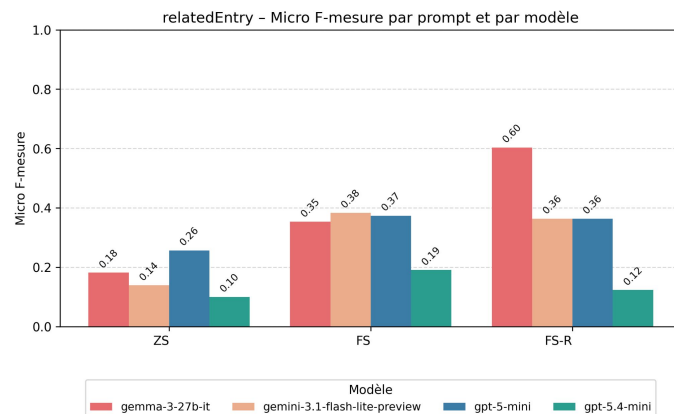
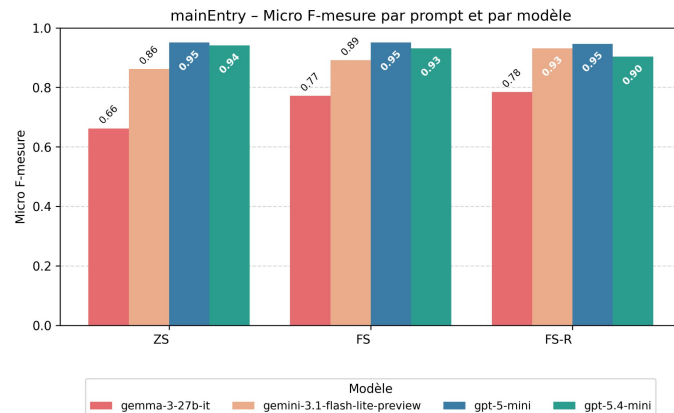


Micro f-mesure par stratégie et modèle pour les entrées et sous-entrées

Prompt unique



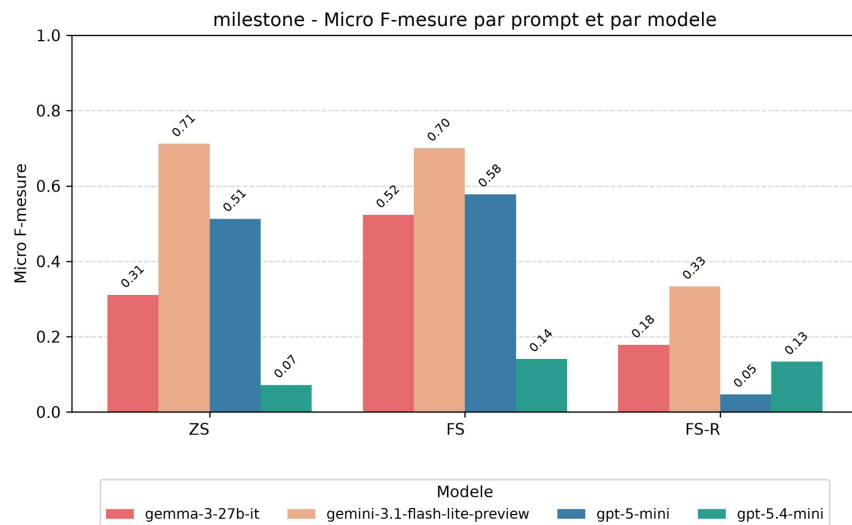
Prompt chaining



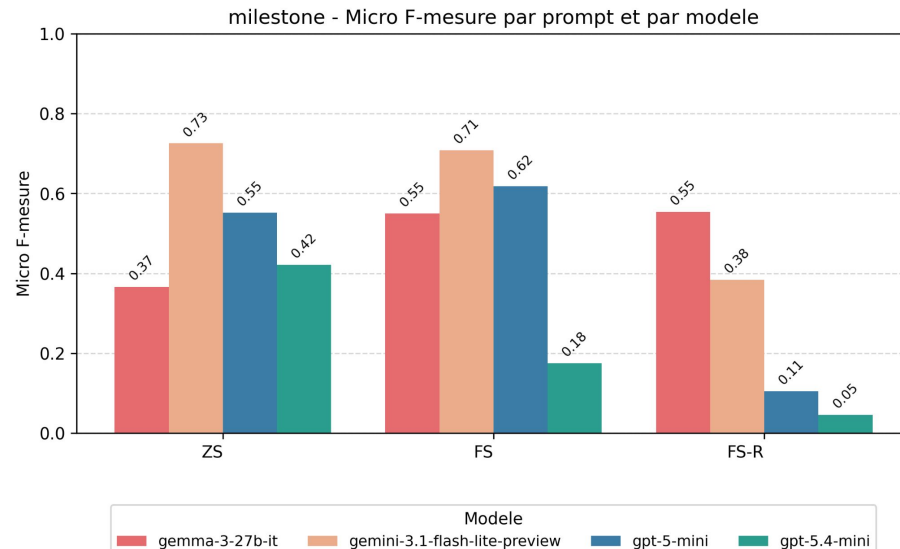
Micro f-mesure par stratégie et modèle pour la balise <milestone>

(bonnes prédictions de gemini
et amélioration de gpt-5-mini grâce au *prompt chaining*)

Prompt unique



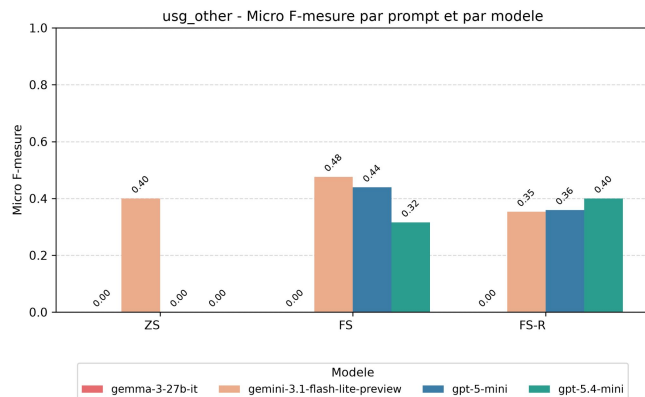
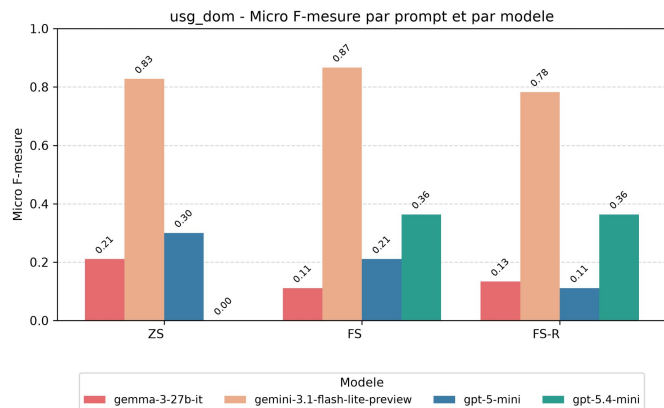
Prompt chaining



Micro f-mesure par stratégie et modèle pour la balise <usg>

(sur l'efficacité du *prompt chaining* pour gpt-5-mini)

Prompt unique



Prompt chaining

