Aspects of comic verse salient to non-specialized large language models: says who?

Pablo Ruiz Fabo
Université de Strasbourg · Universidade de Santiago de Compostela

Plotting Poetry 8 · Prague

Université de Strasbourg

lilpa
linguistique, langues, parole

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CiTIUS
Singular Research Center on
Intelligent technologies

Which aspects of comic verse are salient to a widely used LLM when given very generic instructions?

Which aspects of comic verse are salient to a widely used LLM when given very generic instructions?

Interest: to get an indication of the representations of humor they promote

Humour and LLM

- **JOKER** series of (un)shared tasks at **CLEF** (e.g. Ermakova et al., 2023; 2024) or tasks at **SemEval** (e.g. Meany et al., 2021; Miller et al., 2017)

Tasks

1. Zero-shot binary classification of a poem as comical or not
2. Poem continuation (to test for memorization by the LLM, cf. D'Souza & Mimno, 2023)
3. Naming an author (ditto)

- Prompts: See following slides
- Hyperparameters: default

Prompts: Binary classification

Below instruction + the poem text *[English translation of Spanish prompt used]*

```
"""Is the following poem comic? Why?

Provide your answer in JSON, with the following
structure:
{
 "judgement": "yes|no|uncertain",
 "reason": "reason for your answer"
}

Answer length should be 200 words.
"""
```

Prompts: Poem continuation

Below instruction + the poem text *[English translation of Spanish prompt used]*

```
"""Do you know how the following poem continues?

Provide your answer in JSON, with the following
structure:
{
 "judgement": "yes|no",
 "continuation": "continuation for the poem"
}
"""
```

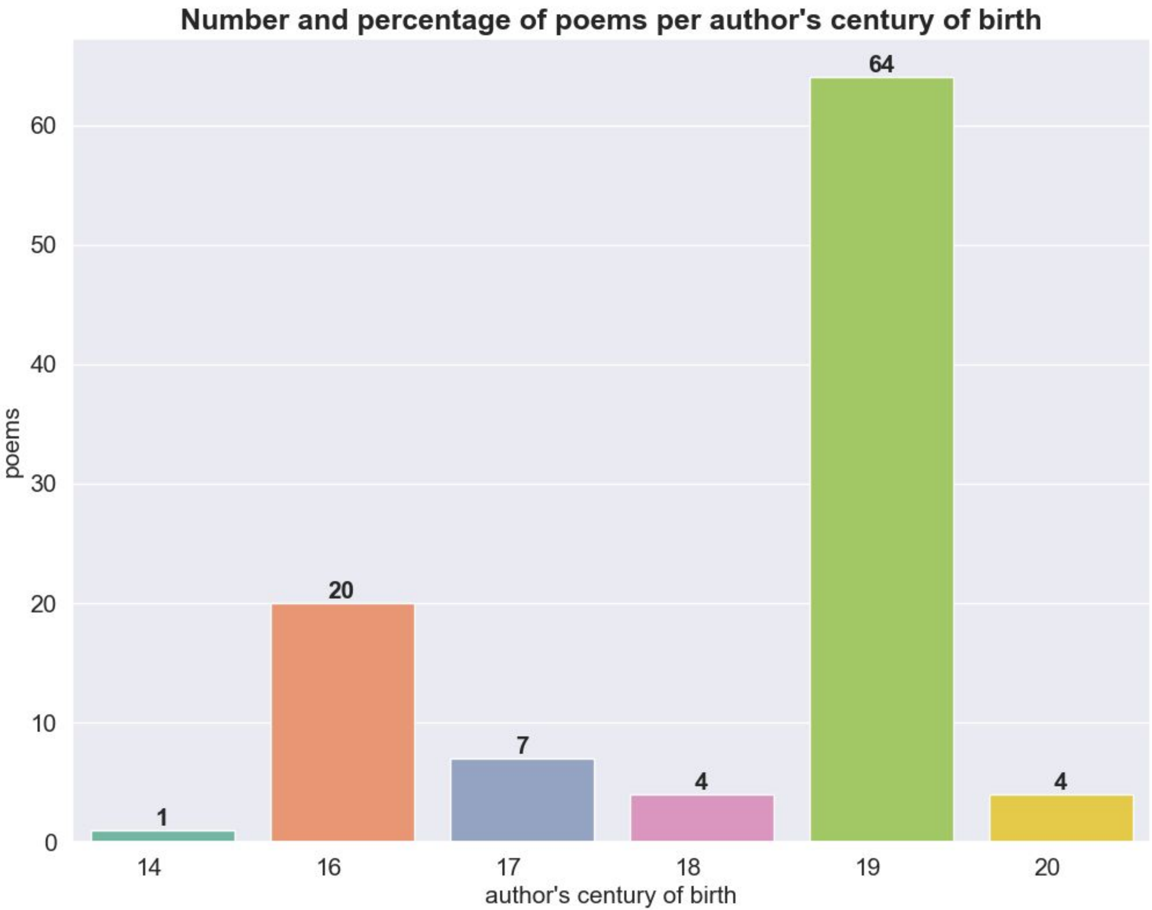
Prompts: Author naming

Below instruction + the poem text *[English translation of Spanish prompt used]*

```
"""Do you know the author for the following poem?
Provide your answer in JSON, with the following
structure:
{
 "author": "first and last name for the author"
 "century": "century when they lived, using arabic
numerals"
}
"""
```

Corpus for prompts

- 100 sonnets in Spanish
  - Comic: 50
  - Serious: 50
- 68 authors
- 15th to 21st century
- Most are not well known



Corpus for prompts: sources

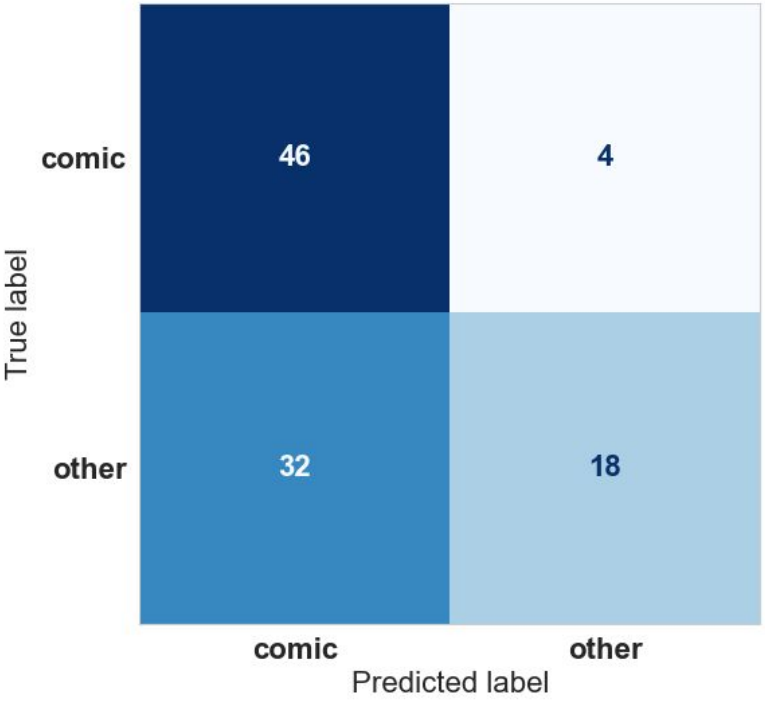
- **Comic part**
  - Some comic sonnet collections (by minor authors)
  - A handful of well-known comical sonnets
  - A handful of sonnets with comical elements but not too clearly humorous
- **Serious part**
  - DISCO corpus (prf1.org/disco) mostly

Corpus for analyses

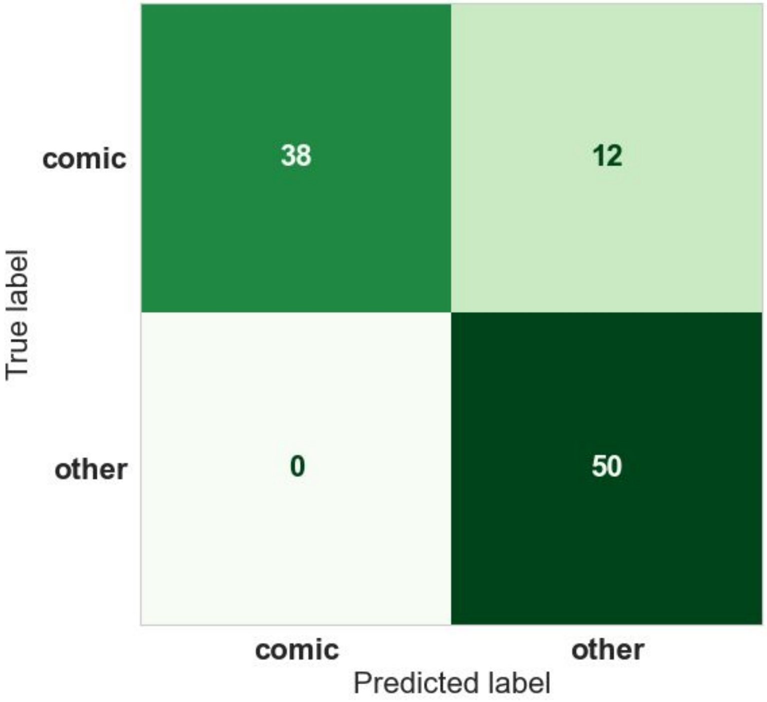
- Humour **judgements**
  - 5 judgements per poem
    - Explicit answer: yes, no, uncertain
    - An explanation for the judgement (asking for 200 words)
- Poem **continuations**
  - Once per poem
  - Explicit answer whether it knows the continuation or not
  - The proposed continuation
- **Author** naming
  - Once per poem
  - Author and its century

Results: Binary classification task

gpt-**3.5**-turbo: **0.61 F1**

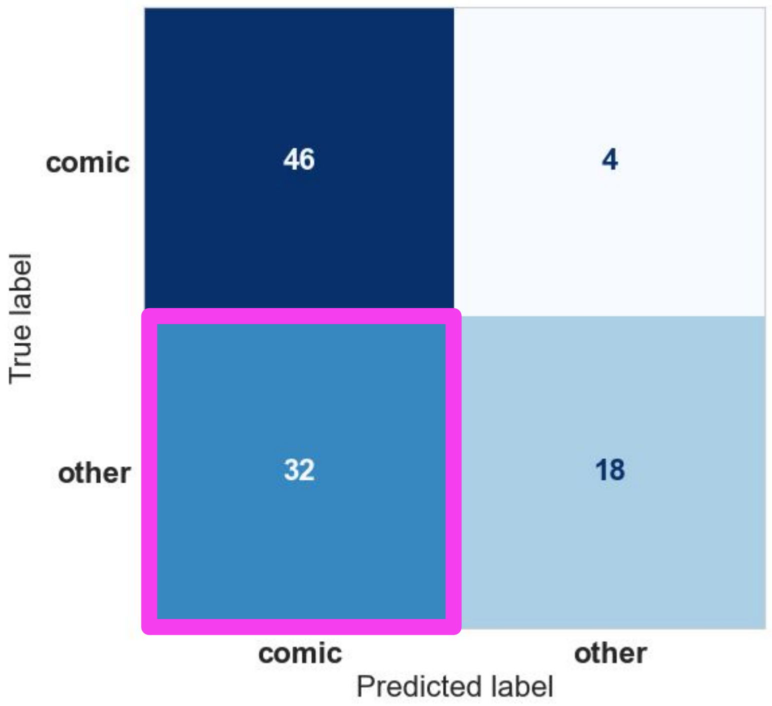


gpt-**4o**: **0.88 F1**

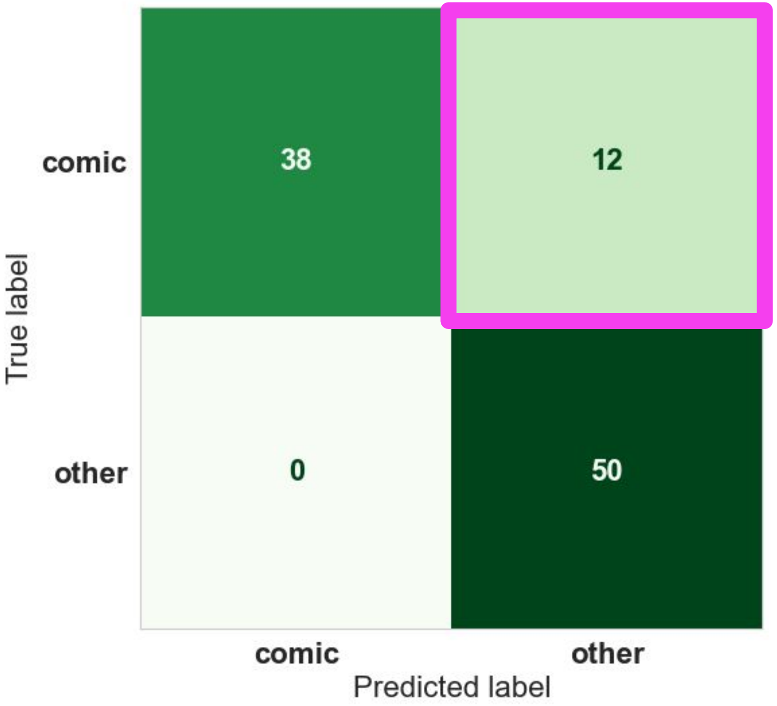


Results: Binary classification task

gpt-**3.5**-turbo: **0.61** **F1**



gpt-**4o**: **0.88** **F1**



Results: Reasons for false positives

- Very common with gpt-**3.5**-turbo (**32 cases**)
- Corpus contains **“classical” sonnets** (i.e. can be unoriginal, conventionally using poetic diction), many from the 19th century, some earlier
- Generated judgement **describes** poem **as** a parody, or language and content is described as **pompous**, ridiculous, **overly dramatic**, absurd
- **Generated text oblivious of poetic conventions in “classical” poetry**
- Generated text fails to show “suspension of poetic disbelief”

century	nbr of fp
16th	6
17th	5
18th	1
19th	19
20th	1

Results: Reasons for false negatives

- More common type of error (**12 cases**) with gpt-**4o**
- Opposite trend as with false positives
- Generated judgement shows no recognition of
    - **subtle irony**
    - **a final punchline**
- Focusing instead on emotional or serious language outside the more ironical parts

Results: poem continuation

- Only one out of the 100 poems was identified, Quevedo's comical sonnet "*A una nariz*" ("To a nose")
- Archi-famous, archi-anthologized

Results: author and century identification

- Only the author for that same poem was identified
- But the better model (4o) tends to provide a closer century to the correct one

Model	**Century matches**	**Mean difference (std) |real - predicted|**	**Correlation (Spearman *r*)**
gpt-3.5-turbo	41	0.8 (0.8)	0.703
gpt-4o	52	0.57 (0.66)	0.779

Results: distinctiveness analysis

- Specificity (Lafon, 1982) and Zeta (Craig, 2008)
- Humour explanations generated correspond to commonly accepted views of humour
  - Schema opposition in comic verse
  - Emotion expression and solemnity as a correlate of non-comic poems

😋

schema opposition
unexpected
contrast
expectation
generate
twist
break
final/ending
touch

rhetorical devices
mockery
ridiculous
absurd
caricature
ingeniousness
mordacity
mixture, combination
extravagance
situation
criticism

(English translations of Spanish terms)

~~😋~~

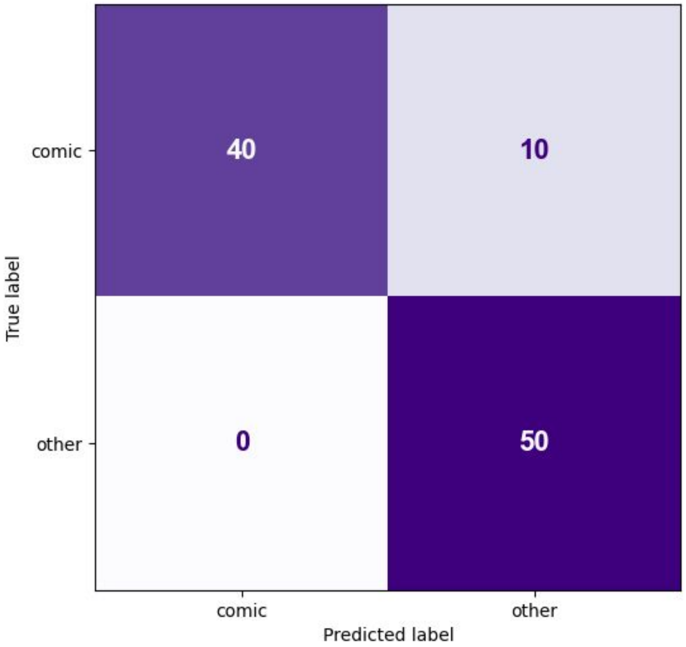
feeling
love
suffering
feeling
emotional
melancholy
sadness
despair
anguish
pain
emotion
passion
desolation
hope

mixed cognition/feeling
solemn
admiration
contemplation
reverence, reverent
homage
reflective
struggle
divine
empathy
perception

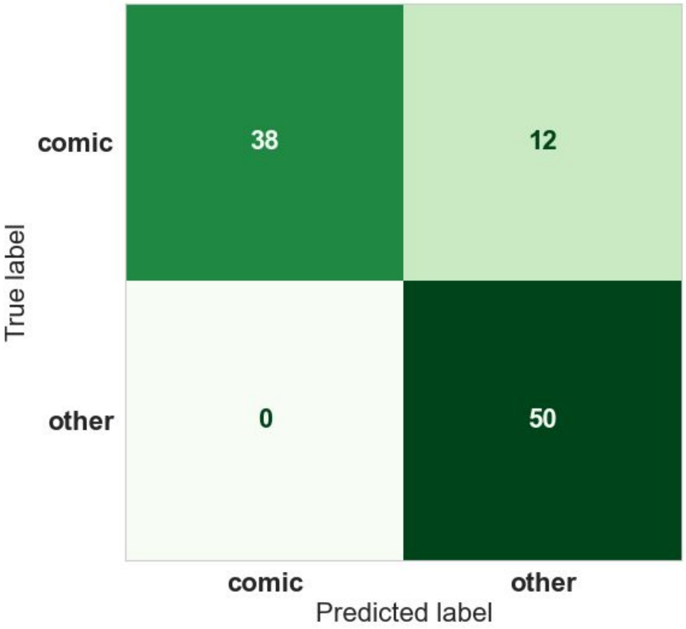
Short-lived generalizations given gpt-4o-mini

- Replaced 3.5-turbo in 07/2024. Very similar results to gpt-4o

gpt-**4o-mini**: **0.90 F1**



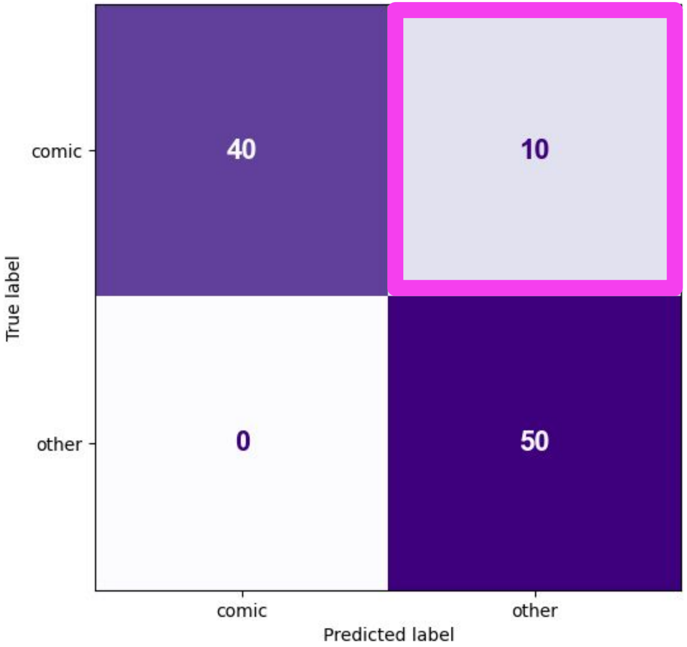
gpt-**4o**: **0.88 F1**



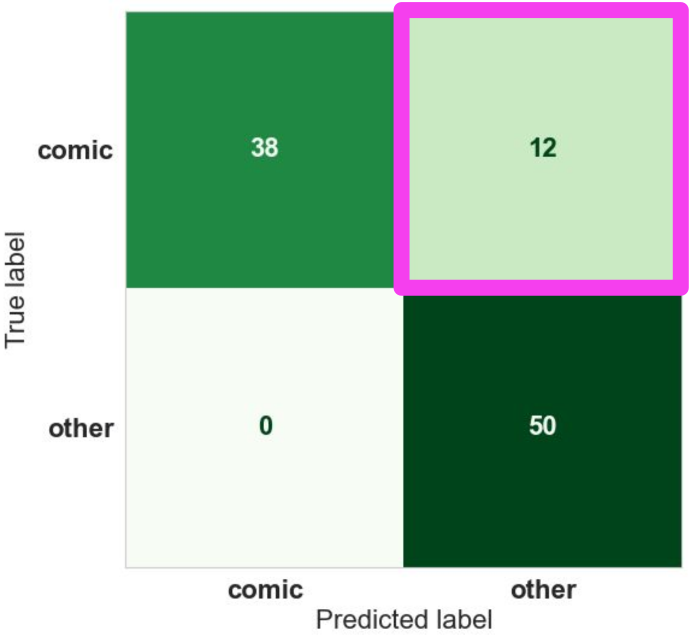
Short-lived generalizations given gpt-4o-mini

- Replaced 3.5-turbo in 07/2024. Very similar results to gpt-4o

gpt-**4o-mini**: **0.90 F1**



gpt-**4o**: **0.88 F1**



New experiments

gpt-3.5-turbo	24-01
gpt-4o-mini	24-07
gpt-4o	24-11
gpt-4.1	25-04

claude-3.5-haiku	24-10
claude-3.5-sonnet	24-10

gemini-1.5-pro	24-10
gemini-2.0-flash	25-02

mistral-large	24-11
mistral-small	25-03
mistral-medium	25-05

deepseek-chat	25-03

New experiments

gpt-3.5-turbo	24-01
gpt-4o-mini	24-07
gpt-4o	24-11
gpt-4.1	25-04

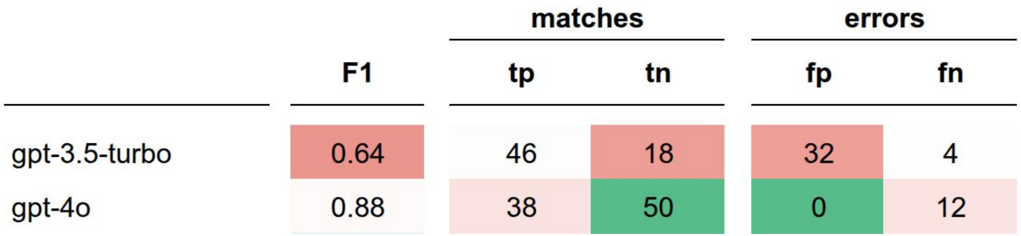
claude-3.5-haiku	24-10
claude-3.5-sonnet	24-10

gemini-1.5-pro	24-10
gemini-2.0-flash	25-02

mistral-large	24-11
mistral-small	25-03
mistral-medium	25-05

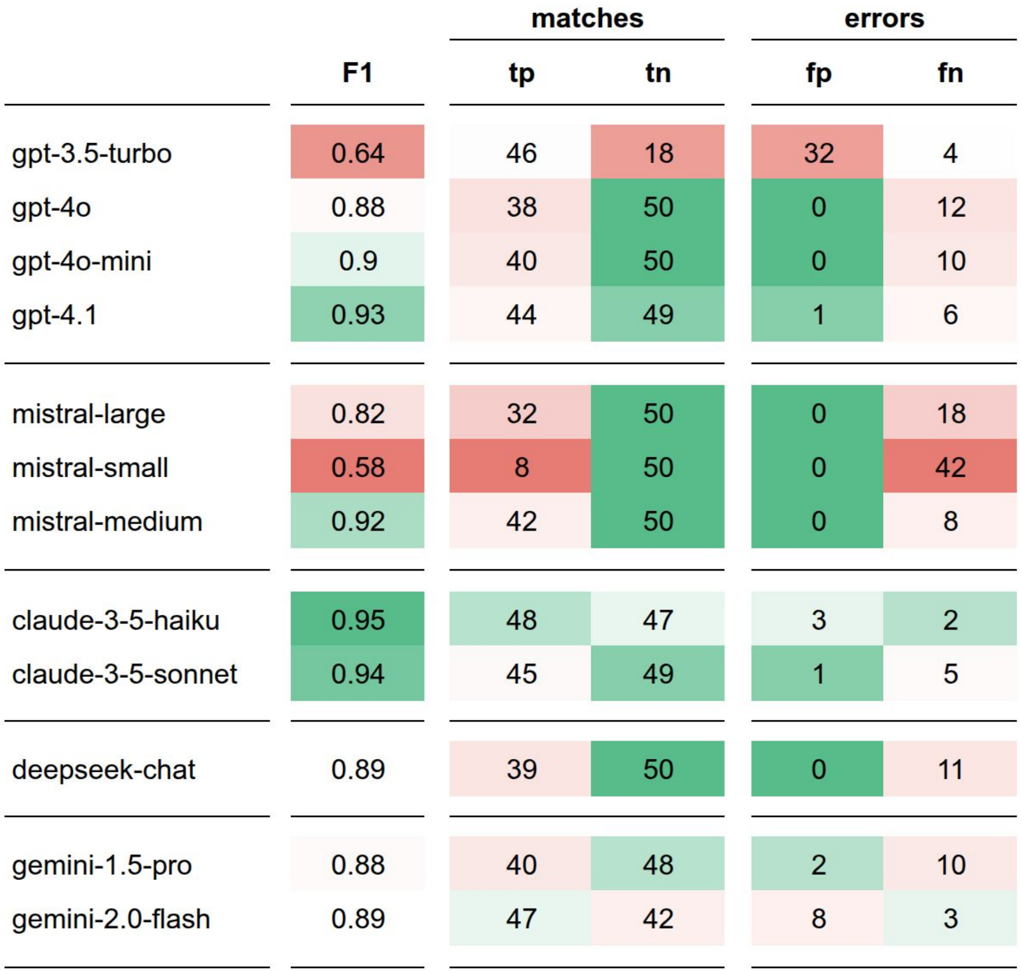
deepseek-chat	25-03

Binary classification task



		matches		errors	
	F1	tp	tn	fp	fn
gpt-3.5-turbo	0.64	46	18	32	4
gpt-4o	0.88	38	50	0	12

Binary classification task



		matches		errors	
	F1	tp	tn	fp	fn
gpt-3.5-turbo	0.64	46	18	32	4
gpt-4o	0.88	38	50	0	12
gpt-4o-mini	0.9	40	50	0	10
gpt-4.1	0.93	44	49	1	6
mistral-large	0.82	32	50	0	18
mistral-small	0.58	8	50	0	42
mistral-medium	0.92	42	50	0	8
claude-3-5-haiku	0.95	48	47	3	2
claude-3-5-sonnet	0.94	45	49	1	5
deepseek-chat	0.89	39	50	0	11
gemini-1.5-pro	0.88	40	48	2	10
gemini-2.0-flash	0.89	47	42	8	3

False negatives

- Generated text **adds nuance** to the task:
  - Poem shows “sarcasm” or “irony”, but is not “comical in a traditional sense”
  - Makes a **perspective** explicit:
    - “may not be humorous for a modern audience” (mistral-medium)
    - “archaic language may cause confusion, not mirth” (mistral-medium)
- Lack of access to **knowledge local** to a tradition
  - Injecting knowledge in prompt may help (RAG, NLP annotations like entities)
  - Adding metadata to prompt (e.g. author origin)

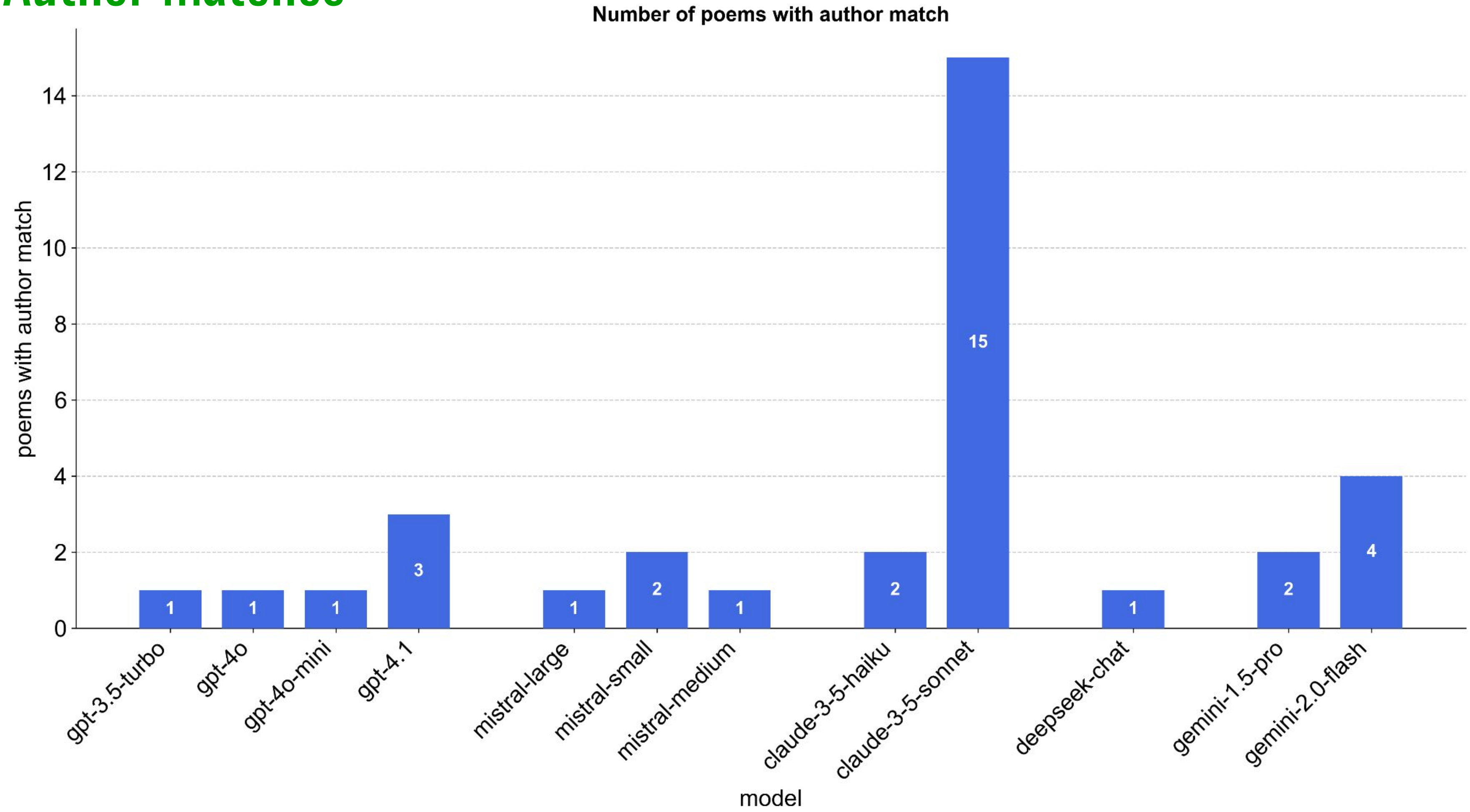
False positives

- **Misrepresentation as parody** or exaggeration, but in some cases:
    - contains correct information about the tradition: “friendly parody of the modernist sonnet” for an *actual* modernist sonnet (claude-sonnet)
    - mentions the “modern perspective” as the reason to consider it comical (gemini-1.5-pro)

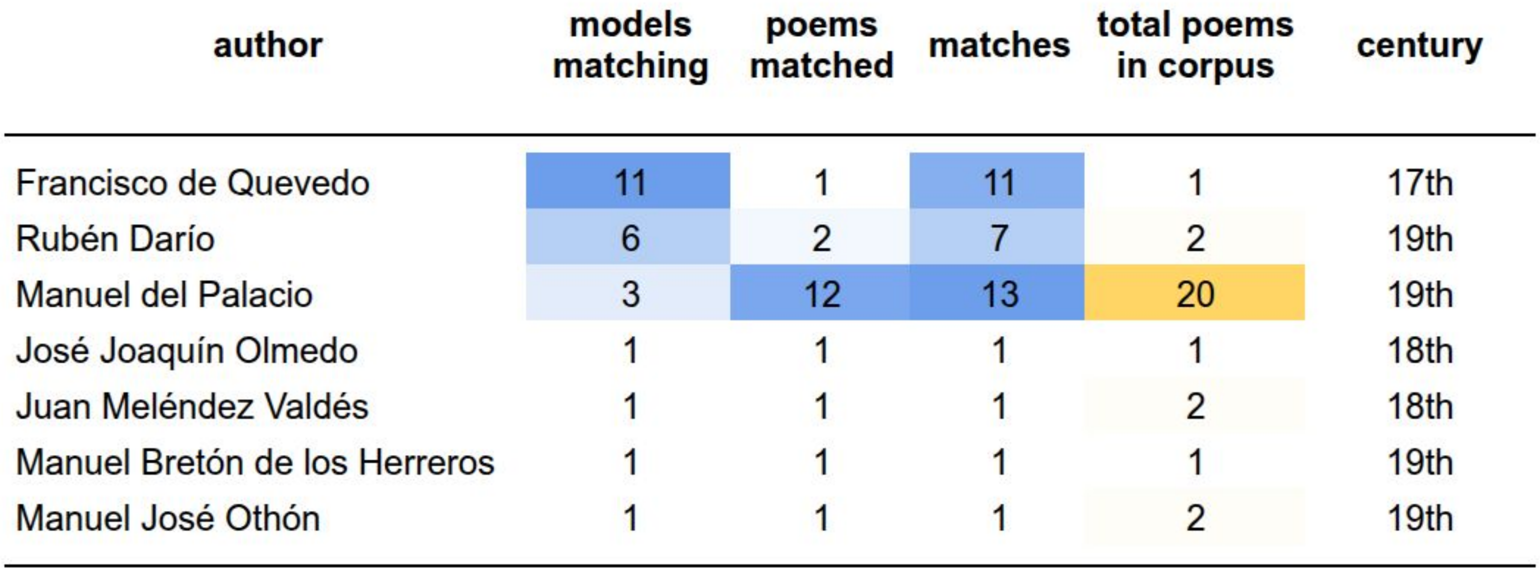
Rhythmic elements in the judgements?

- mistral-small: hallucinated rhyme schemes
    - ABABAB CDCDCD EFEFEF and ABABABCDCDEE (for ABBA sonnets)
- claude-haiku
    - sonnet has “ABBA or cuartet**a** form” → should be cuartet***o***, as > 8 syllables
- claude-sonnet and gemini
    - texts identify alternate and enclosed quatrains coherently (ABAB vs. ABBA)

Author matches

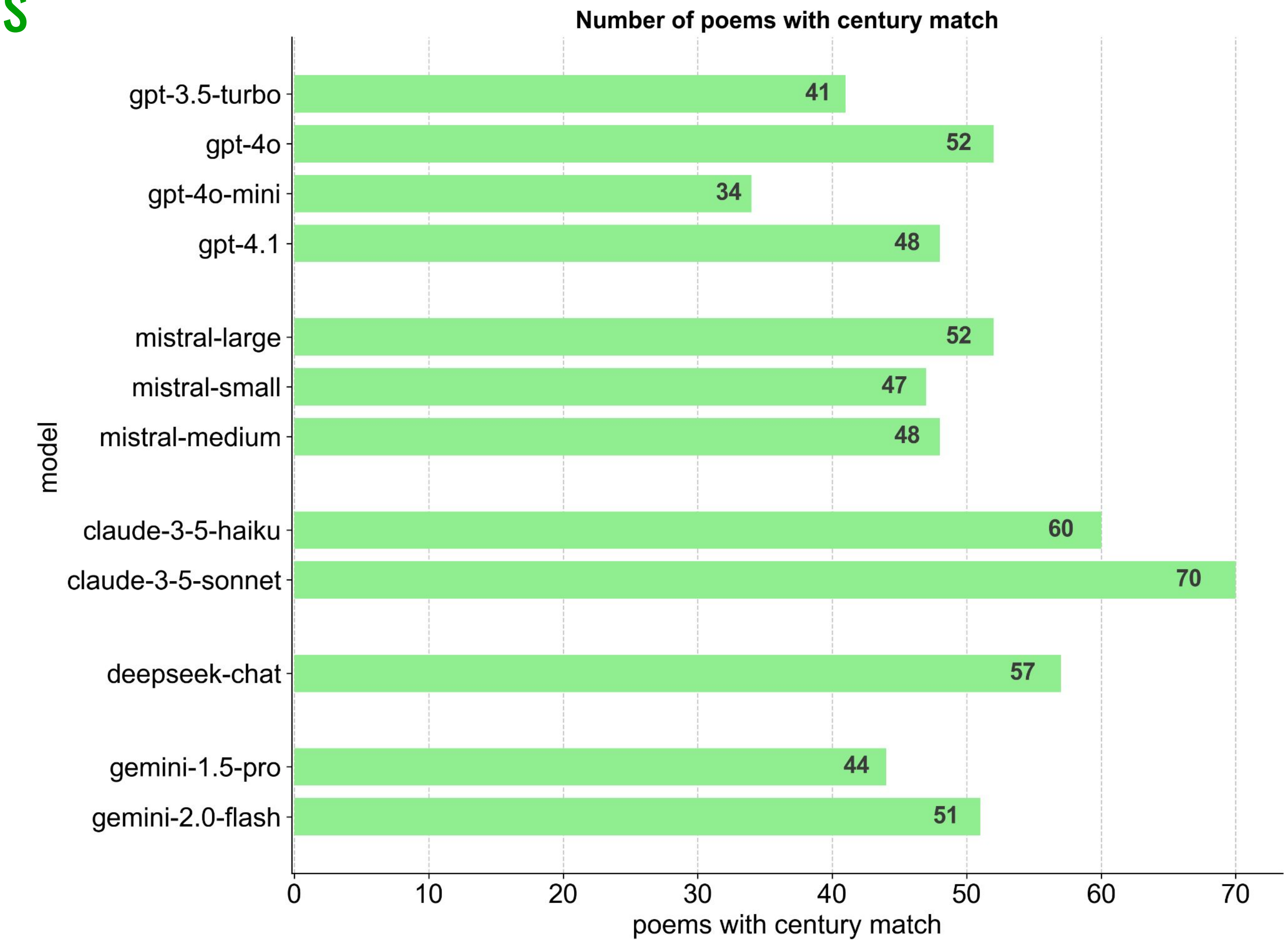


Author matches

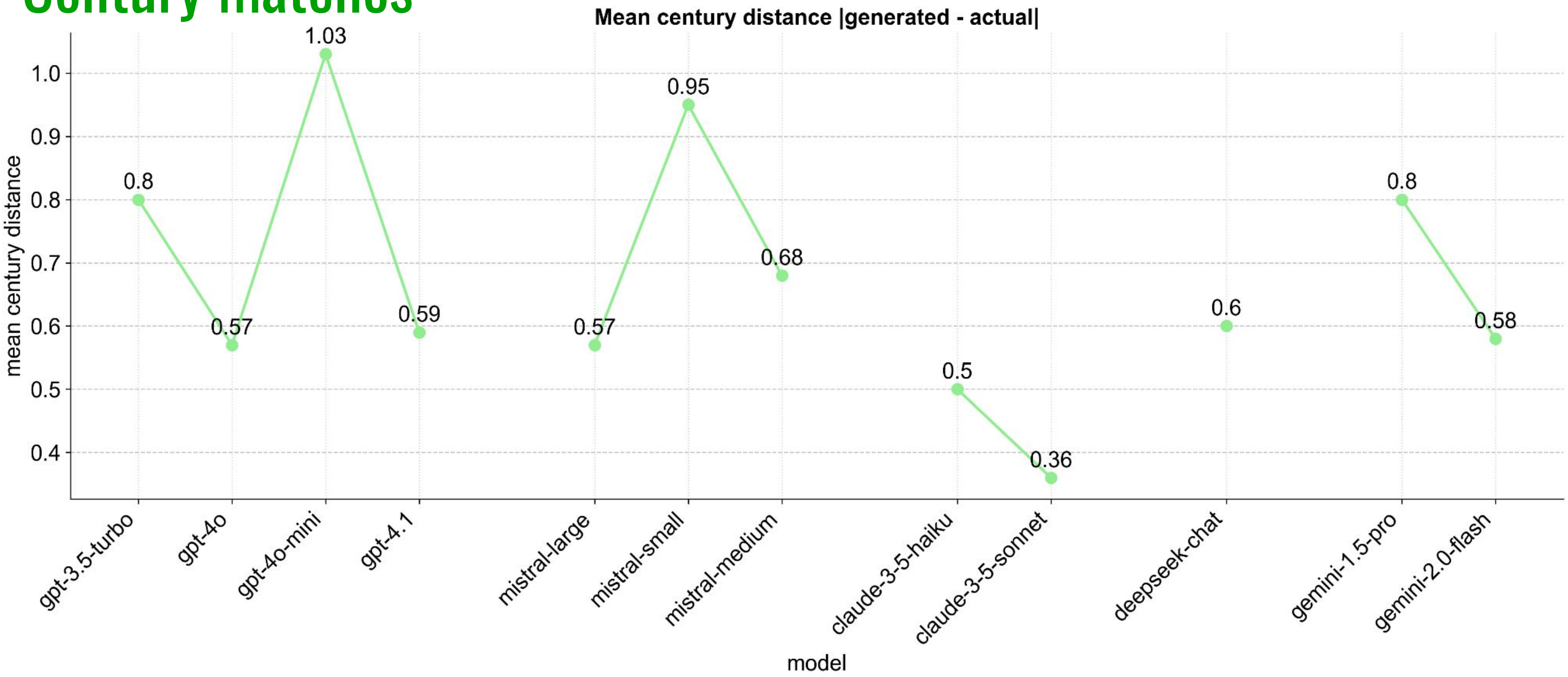


author	models matching	poems matched	matches	total poems in corpus	century
Francisco de Quevedo	11	1	11	1	17th
Rubén Darío	6	2	7	2	19th
Manuel del Palacio	3	12	13	20	19th
José Joaquín Olmedo	1	1	1	1	18th
Juan Meléndez Valdés	1	1	1	2	18th
Manuel Bretón de los Herreros	1	1	1	1	19th
Manuel José Othón	1	1	1	2	19th

Century matches



Century matches



0.72	0.8	0.54	0.79	0.8	0.59	0.7	0.78	0.85	0.74	0.63	0.78

Spearman ρ: actual and generated centuries

Poem continuation

- Only two of the poems are given a correct continuation (for some lines, then diverges)
  - “A una nariz” (Quevedo) + “Melancolía” (Darío)
- In many cases, models assert knowing the continuation, but then generate an unrelated text
  - An occasion to witness their generation skills in classical poetry

To sum up

- Trends in the results with new models tested:
    - False negatives highlight the underspecification of the task (binary *humor/not* classification, whereas text generated adds nuance)
    - False positives sometimes are explicitly accompanied by the mention of “the current perspective”
- Better models show better access to literary history information in that they generate a more accurate century and slightly more matching authors
- Claude family the most poetry friendly according to this task and dataset

Model identification based on the humour explanations

- Trained classical models for supervised classification
- Features: tf-idf weighted word-form unigrams and PoS unigrams
    - with and without punctuation
- Models: Logistic Regression, Linear SVM (via `scikit-learn`)
- Data: 500 generated texts (ca 10K total tokens) per model, 80/20 split

Model recognition based on the humour explanations

Logistic Regression

```
                           precision    recall  f1-score   support

   claude-3-5-haiku-latest    0.9794    0.9500    0.9645       100
  claude-3-5-sonnet-latest    0.9412    0.9600    0.9505       100
             deepseek-chat    0.9200    0.9200    0.9200       100
            gemini-15-pro    0.8393    0.9400    0.8868       100
          gemini-20-flash    0.9200    0.9200    0.9200       100
             gpt-35-turbo    0.8571    0.9000    0.8780       100
                   gpt-41    0.9000    0.9000    0.9000       100
                   gpt-4o    0.7561    0.6200    0.6813       100
              gpt-4o-mini    0.9691    0.9400    0.9543       100
     mistral-large-latest    0.8571    0.8400    0.8485       100
           mistral-medium    0.9126    0.9400    0.9261       100
            mistral-small    0.9038    0.9400    0.9216       100

                 accuracy                        0.8975      1200
                macro avg    0.8963    0.8975    0.8960      1200
             weighted avg    0.8963    0.8975    0.8960      1200
```

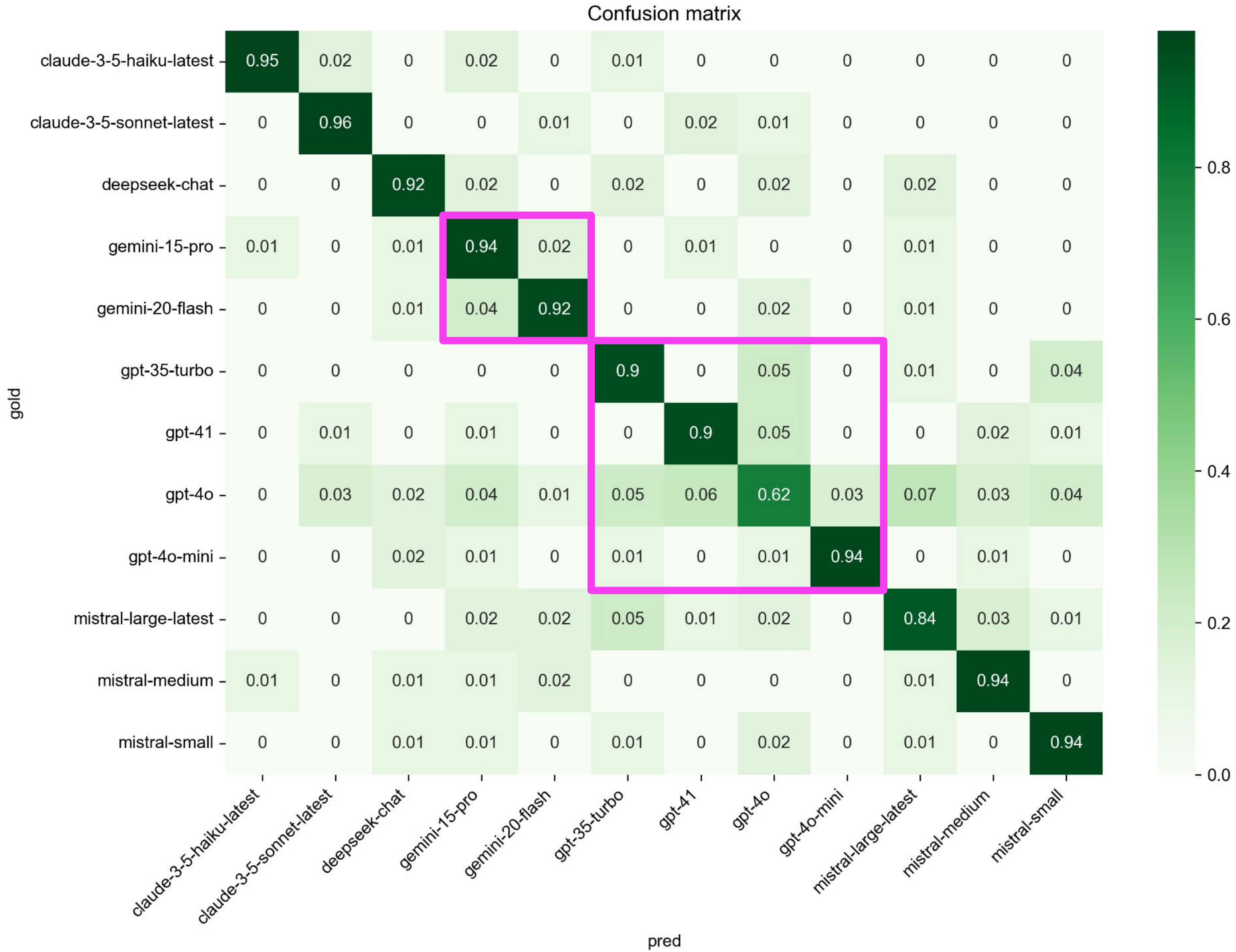
Model recognition based on the humour explanations

Linear SVM

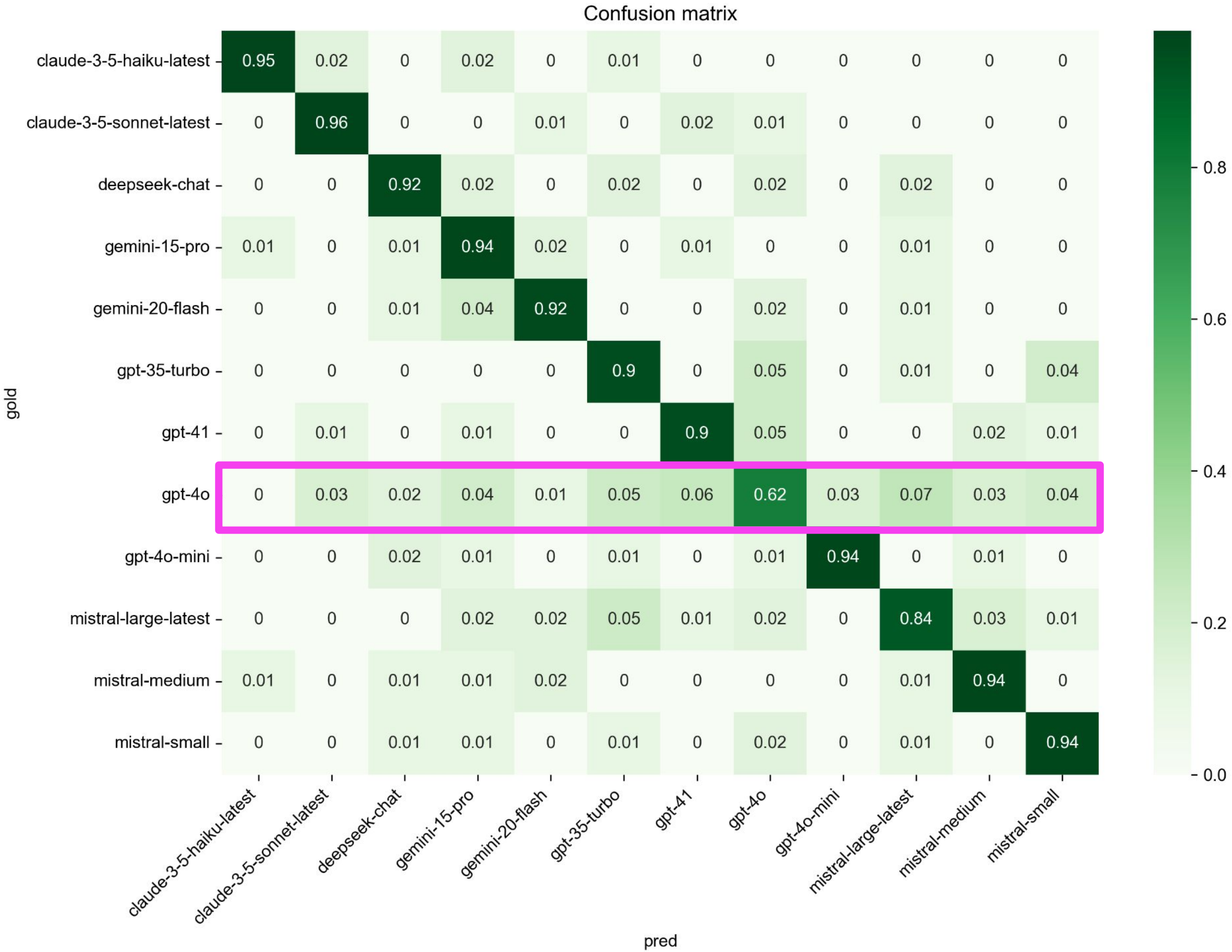
```
                          precision    recall  f1-score   support

  claude-3-5-haiku-latest    0.9412    0.9600    0.9505       100
 claude-3-5-sonnet-latest    0.9307    0.9400    0.9353       100
            deepseek-chat    0.9348    0.8600    0.8958       100
            gemini-15-pro    0.7826    0.9000    0.8372       100
          gemini-20-flash    0.8641    0.8900    0.8768       100
             gpt-35-turbo    0.8400    0.8400    0.8400       100
                   gpt-41    0.9091    0.9000    0.9045       100
                   gpt-4o    0.7662    0.5900    0.6667       100
              gpt-4o-mini    0.9515    0.9800    0.9655       100
     mistral-large-latest    0.8200    0.8200    0.8200       100
           mistral-medium    0.9126    0.9400    0.9261       100
            mistral-small    0.8667    0.9100    0.8878       100

                 accuracy                        0.8775      1200
                macro avg    0.8766    0.8775    0.8755      1200
             weighted avg    0.8766    0.8775    0.8755      1200
```



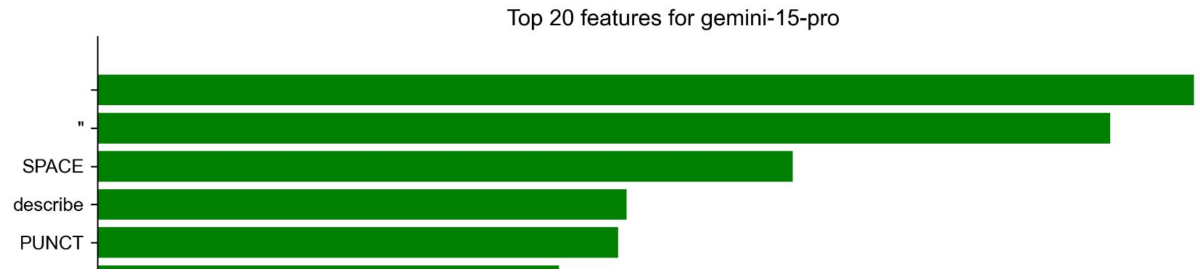
gold \ pred	claude-3-5-haiku-latest	claude-3-5-sonnet-latest	deepseek-chat	gemini-15-pro	gemini-20-flash	gpt-35-turbo	gpt-41	gpt-4o	gpt-4o-mini	mistral-large-latest	mistral-medium	mistral-small
claude-3-5-haiku-latest	0.95	0.02	0	0.02	0	0.01	0	0	0	0	0	0
claude-3-5-sonnet-latest	0	0.96	0	0	0.01	0	0.02	0.01	0	0	0	0
deepseek-chat	0	0	0.92	0.02	0	0.02	0	0.02	0	0.02	0	0
gemini-15-pro	0.01	0	0.01	0.94	0.02	0	0.01	0	0	0.01	0	0
gemini-20-flash	0	0	0.01	0.04	0.92	0	0	0.02	0	0.01	0	0
gpt-35-turbo	0	0	0	0	0	0.9	0	0.05	0	0.01	0	0.04
gpt-41	0	0.01	0	0.01	0	0	0.9	0.05	0	0	0.02	0.01
gpt-4o	0	0.03	0.02	0.04	0.01	0.05	0.06	0.62	0.03	0.07	0.03	0.04
gpt-4o-mini	0	0	0.02	0.01	0	0.01	0	0.01	0.94	0	0.01	0
mistral-large-latest	0	0	0	0.02	0.02	0.05	0.01	0.02	0	0.84	0.03	0.01
mistral-medium	0.01	0	0.01	0.01	0.02	0	0	0	0	0.01	0.94	0
mistral-small	0	0	0.01	0.01	0	0.01	0	0.02	0	0.01	0	0.94



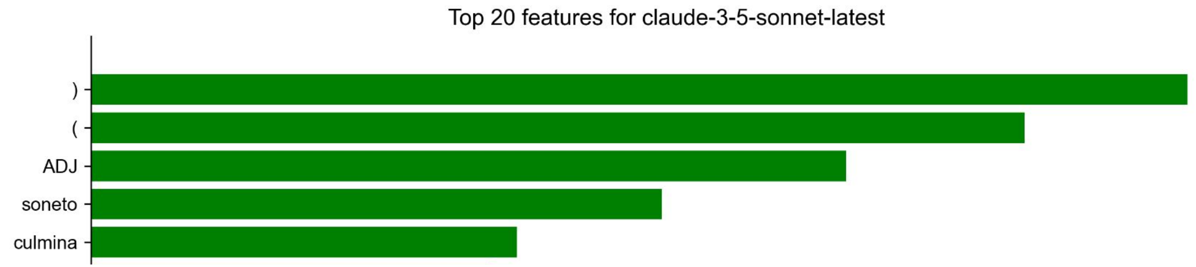
Confusion matrix

gold \ pred	claude-3-5-haiku-latest	claude-3-5-sonnet-latest	deepseek-chat	gemini-15-pro	gemini-20-flash	gpt-35-turbo	gpt-41	gpt-4o	gpt-4o-mini	mistral-large-latest	mistral-medium	mistral-small
claude-3-5-haiku-latest	0.95	0.02	0	0.02	0	0.01	0	0	0	0	0	0
claude-3-5-sonnet-latest	0	0.96	0	0	0.01	0	0.02	0.01	0	0	0	0
deepseek-chat	0	0	0.92	0.02	0	0.02	0	0.02	0	0.02	0	0
gemini-15-pro	0.01	0	0.01	0.94	0.02	0	0.01	0	0	0.01	0	0
gemini-20-flash	0	0	0.01	0.04	0.92	0	0	0.02	0	0.01	0	0
gpt-35-turbo	0	0	0	0	0	0.9	0	0.05	0	0.01	0	0.04
gpt-41	0	0.01	0	0.01	0	0	0.9	0.05	0	0	0.02	0.01
gpt-4o	0	0.03	0.02	0.04	0.01	0.05	0.06	0.62	0.03	0.07	0.03	0.04
gpt-4o-mini	0	0	0.02	0.01	0	0.01	0	0.01	0.94	0	0.01	0
mistral-large-latest	0	0	0	0.02	0.02	0.05	0.01	0.02	0	0.84	0.03	0.01
mistral-medium	0.01	0	0.01	0.01	0.02	0	0	0	0	0.01	0.94	0
mistral-small	0	0	0.01	0.01	0	0.01	0	0.02	0	0.01	0	0.94

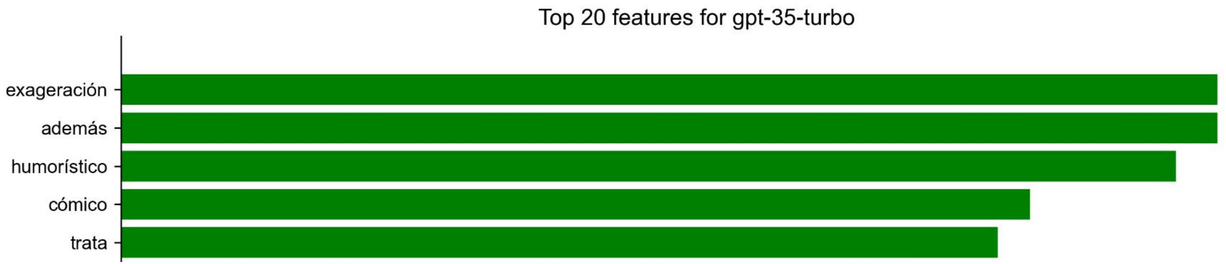
Logistic Regression: Feature importances



Top 20 features for gemini-15-pro

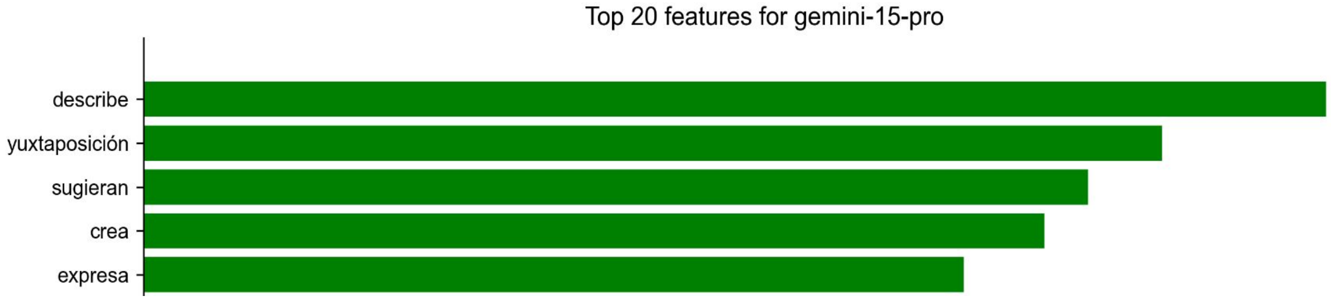


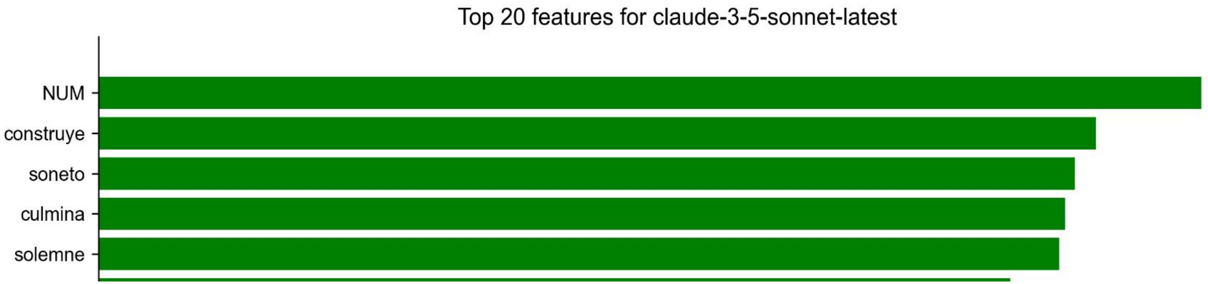
Top 20 features for claude-3-5-sonnet-latest

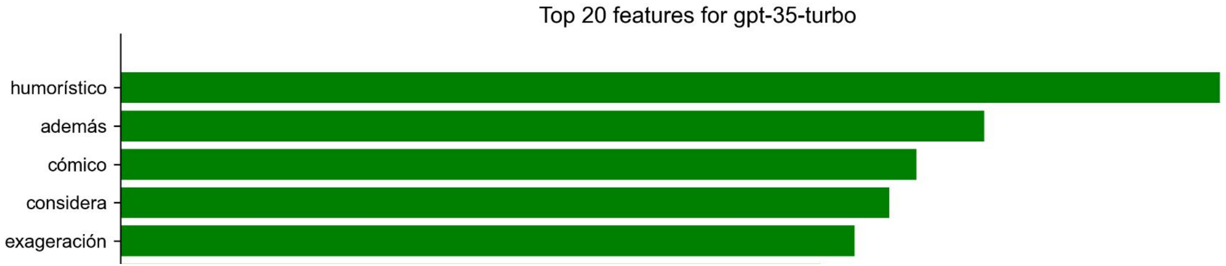


Top 20 features for gpt-35-turbo

Linear SVM: Feature importances


Top 20 features for gemini-15-pro


Top 20 features for claude-3-5-sonnet-latest


Top 20 features for gpt-35-turbo

Bibliography

Brunelière, O., Germann, C., & Salina, K. (2023). *CLEF 2023 JOKER Task 2 : Using ChatGPT For Pun Location And Interpretation.*

D'Souza, L., & Mimno, D. (2023). The Chatbot and the Canon: Poetry Memorization in LLMs. *Computational Humanities Research*. https://ceur-ws.org/Vol-3558/paper5712.pdf

Craig, H., & Kinney, A. F. (Éds.). (2009). *Shakespeare, Computers, and the Mystery of Authorship* (1st éd.). Cambridge University Press.

Ermakova, L., Bosser, A.-G., Miller, T., Thomas, T., Preciado, V. M. P., Sidorov, G., & Jatowt, A. (2024). CLEF 2024 JOKER Lab : Automatic Humour Analysis. In N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, & I. Ounis (Éds.), *Advances in Information Retrieval* (p. 36‑43). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-56072-9_5

Ermakova, L., Miller, T., Bosser, A.-G., Palma Preciado, V. M., Sidorov, G., & Jatowt, A. (2023). Overview of JOKER – CLEF-2023 Track on Automatic Wordplay Analysis. In A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, & N. Ferro (Éds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction* (Vol. 14163, p. 397‑415). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-42448-9_26

Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, *1*(1), 127‑165. https://doi.org/10.3406/mots.1980.1008

Bibliography

Meaney, J. A., Wilson, S., Chiruzzo, L., Lopez, A., & Magdy, W. (2021). SemEval 2021 Task 7 : HaHackathon, Detecting and Rating Humor and Offense. In A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, & X. Zhu (Éds.), *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)* (p. 105‑119). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.semeval-1.9

Miller, T., Hempelmann, C., & Gurevych, I. (2017). SemEval-2017 Task 7 : Detection and Interpretation of English Puns. In S. Bethard, M. Carpuat, M. Apidianaki, S. M. Mohammad, D. Cer, & D. Jurgens (Éds.), *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (p. 58‑68). Association for Computational Linguistics. https://doi.org/10.18653/v1/S17-2005

Popova, O., & Dadić, P. (2023). Does AI Have a Sense of Humor ? CLEF 2023 JOKER Tasks 1, 2 and 3 : Using BLOOM, GPT, SimpleT5, and More for Pun Detection, Location, Interpretation and Translation.

Preciado, V. M. P., Preciado, C. P., & Sidorov, G. (2023). NLPalma @ CLEF 2023 JOKER: A BLOOMZ and BERT Approach for Wordplay Detection and Translation.

Thank you!

Data and code: https://github.com/pruizf/pp8

Prompts: Binary classification

Below instruction + the poem text

```
"""¿Es cómico el poema siguiente? ¿Por qué?

Da una respuesta en JSON con la siguiente estructura:
{
 "judgement": "sí|no|incierto",
 "reason": "razón de la respuesta"
}

La longitud de la respuesta debe ser de 200 palabras.
"""
```

Asked for 5 completions

Prompts: Poem continuation

Below instruction + the poem text

```
"""¿Sabes cómo continúa el poema siguiente?

Da una respuesta en JSON con la siguiente estructura:
{
 "judgement": "sí|no",
 "continuation": "continuación del poema"
}
"""
```

Prompts: Author naming

Below instruction + the poem text

```
"""¿Sabes quién es el autor o autora del poema
siguiente?
Da una respuesta en JSON con la siguiente
estructura:
{
 "author": "nombre y apellidos del autor o autora"
 "century": "siglo en el que vivió, en números
arábigos"
}
"""
```