Can Large Language Models compete with specialized models in Lexical Semantic Change Detection?

Frank D. Zamora-Reina
Felipe Bravo-Márquez
University of Chile, CENIA and IMFD

Dominik Schlechtweg
University of Stuttgart

Nikolay Arefyev
University of Oslo

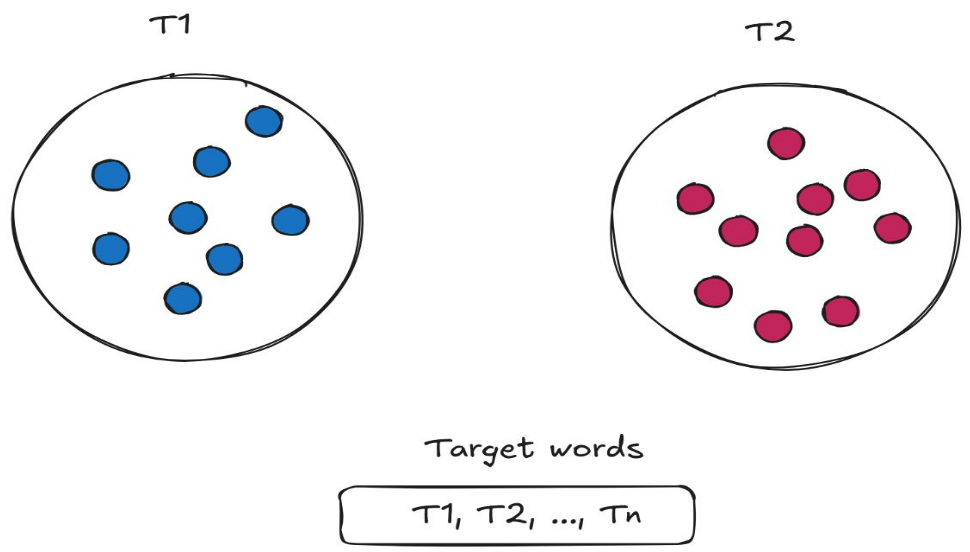
Research questions

- [RQ1] Can automatically optimized prompts yield better results for the LSCD task than manually crafted prompts designed through prompt engineering?

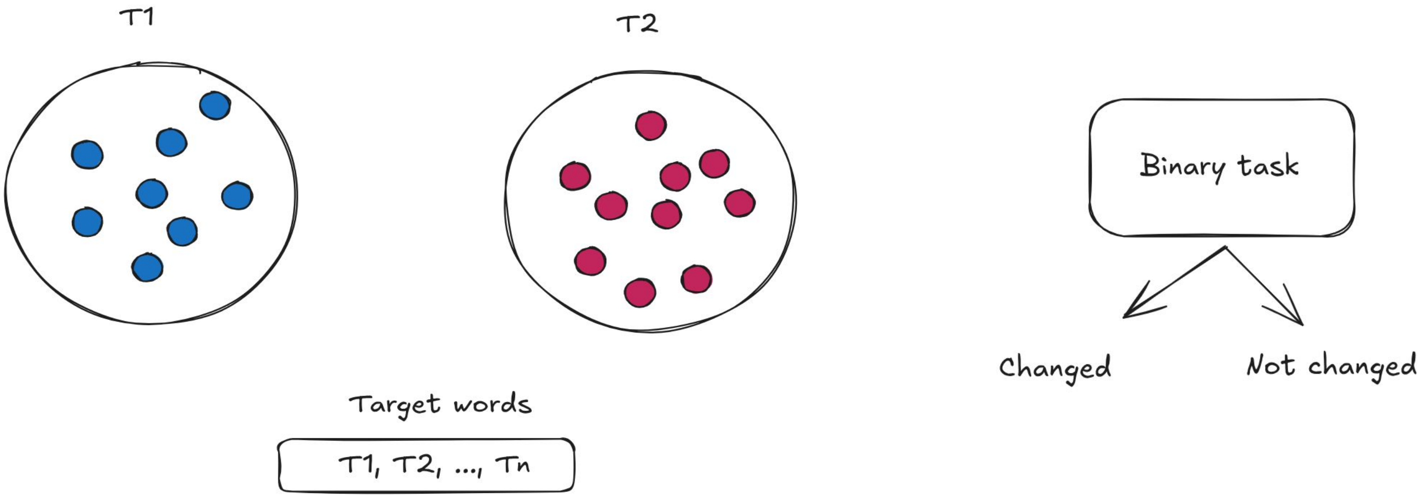
- [RQ2] Can LLMs solve the Graded Change LSCD task well? Can these results surpass the WiC models reported as state-of-the-art?

- [RQ3] Can LLMs outperform state-of-the-art LSCD models at the annotation level?

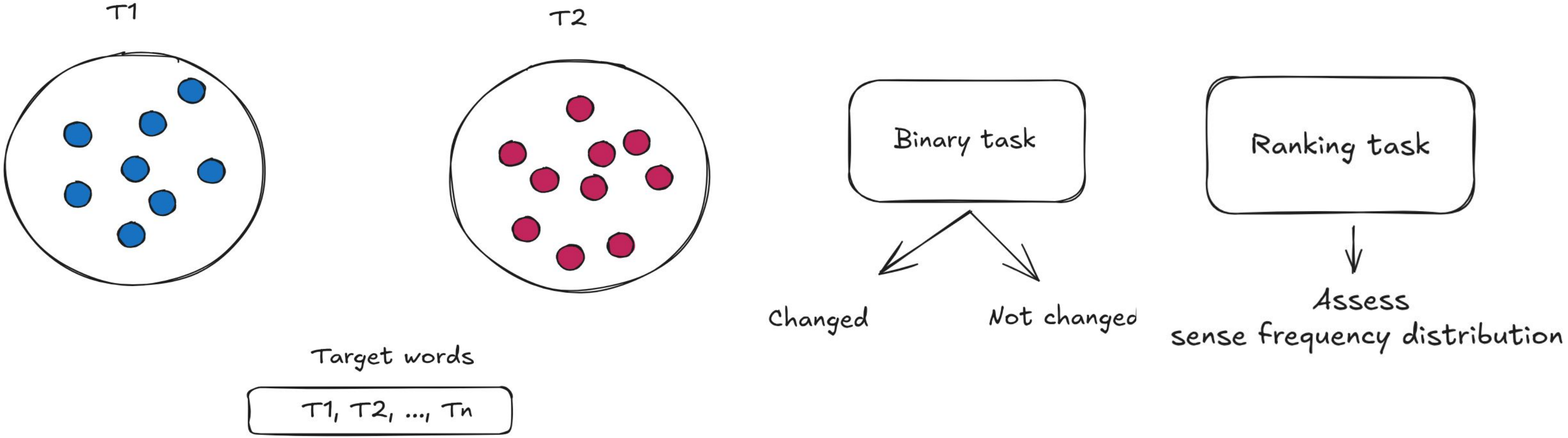
Lexical Semantic Change Detection



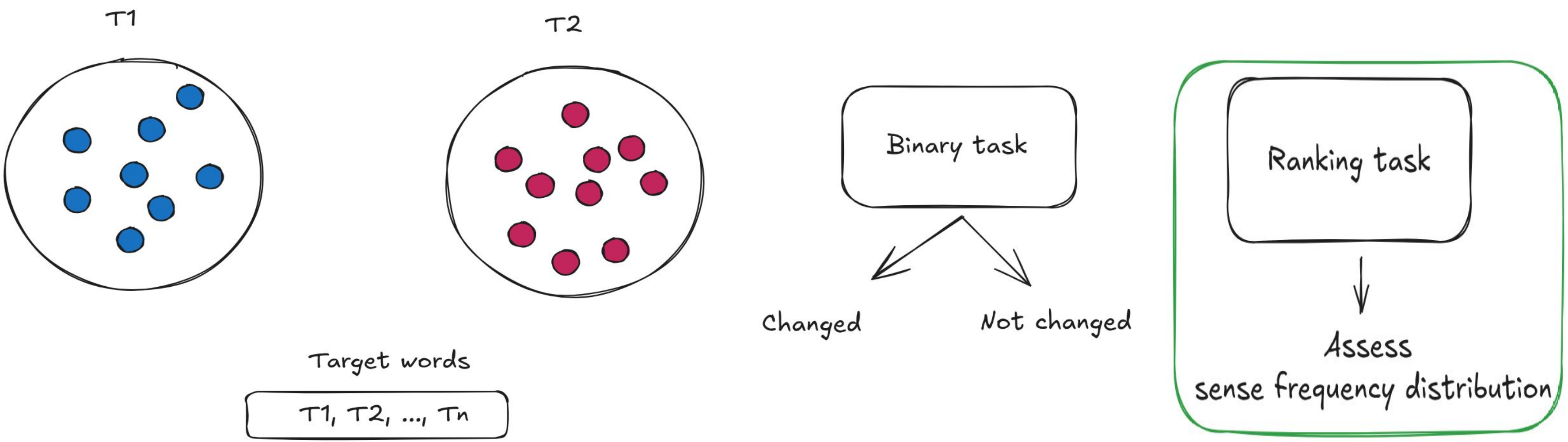
Lexical Semantic Change Detection



Lexical Semantic Change Detection



Lexical Semantic Change Detection



Metrics

- SPR_WiC - Spearman correlation between the annotations provided by human annotators and models.

- SPR_LSCD - Spearman correlation between gold graded change scores and model predictions

Data

- DWUG EN (Schlechtweg et al., 2020)
  - 37 target words (~23k usage pairs)
  - (1810 - 1860) - (1960 - 2010)
- DWUG ES (Zamora-Reina et al., 2022)
  - 60 target words (~27k usage pairs)
  - (1810 - 1906) - (1994 - 2020)

(Schlechtweg et al., 2021)

Data

first usage: His parents had left a lot of money in the bank and now it was Measle's, but a judge had said that Measle was too young to get it.

second usage: Sherrell, it is said, was sitting on the bank of the river close by, and as soon as the men had disappeared from sight he jumped on board the schooner.

target word: bank

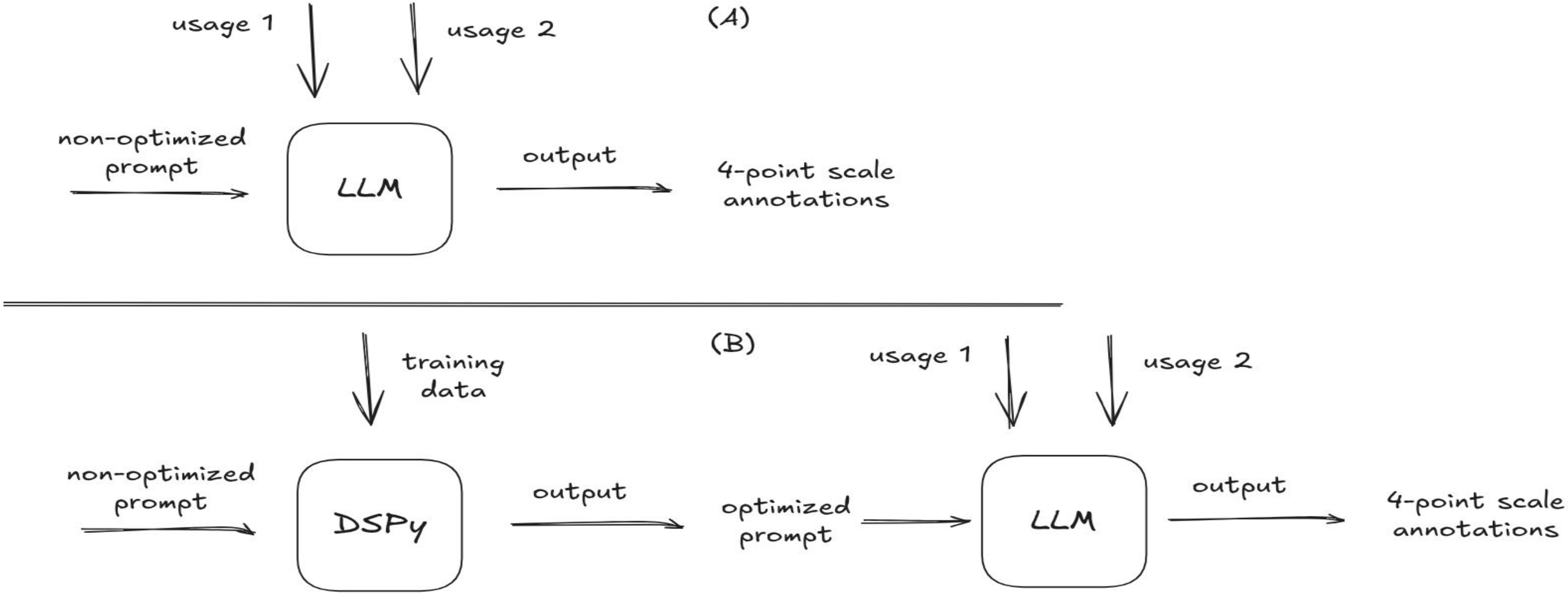
annotation → 4: Identical
3: Closely Related
2: Distantly Related
1: Unrelated

(Schlechtweg et al., 2018)

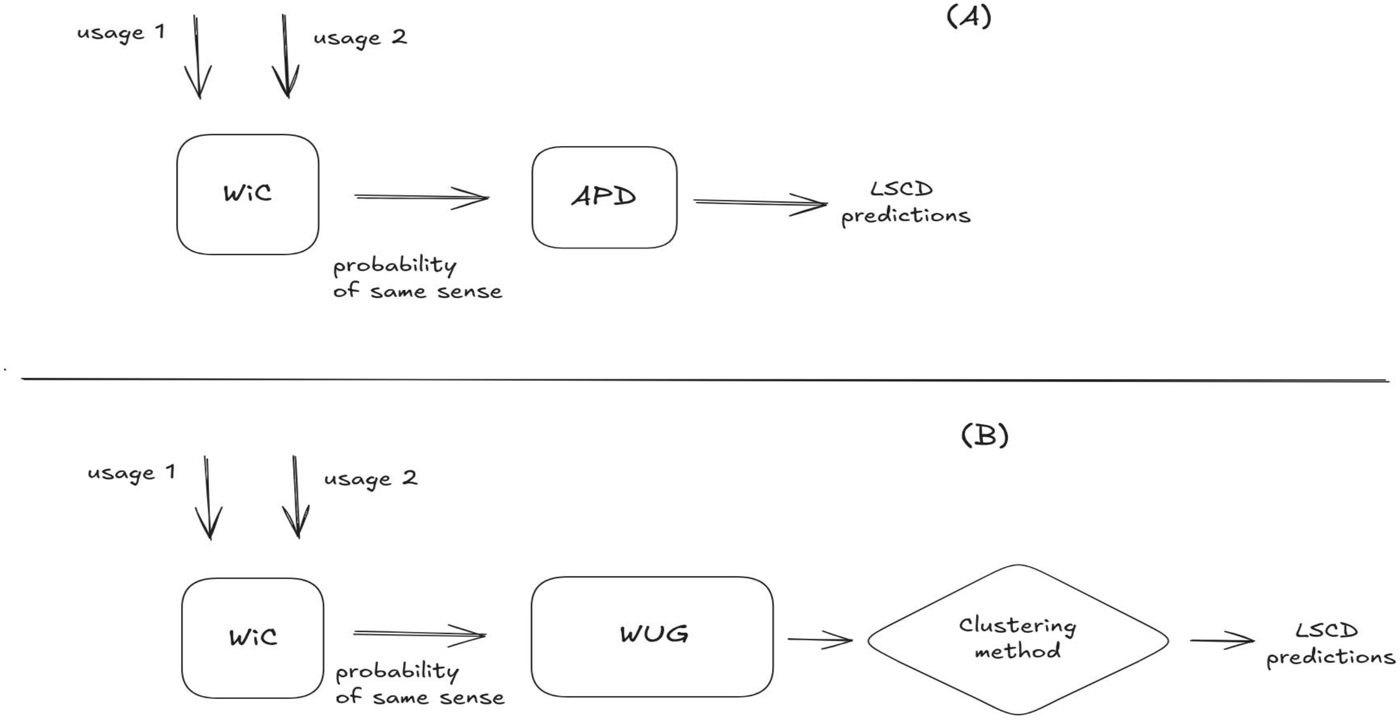
Models

- LLMs
  - Llama 3.1 (8B)
  - Mixtral 8x7B (Jiang et al., 2024)
  - Llama 3.3 (70B)
- WiC models
  - DeepMistake (Arefyev et al., 2021)
    - SOTA for Russian and Spanish datasets
    - MCL, enMCL, MCL-> es (fine-tuned on various WiC datasets across multiple languages)
  - XL-LEXEME (Casotti et al., 2023)
    - SOTA for English, Swedish, German datasets

Experimental Setup - LLMs

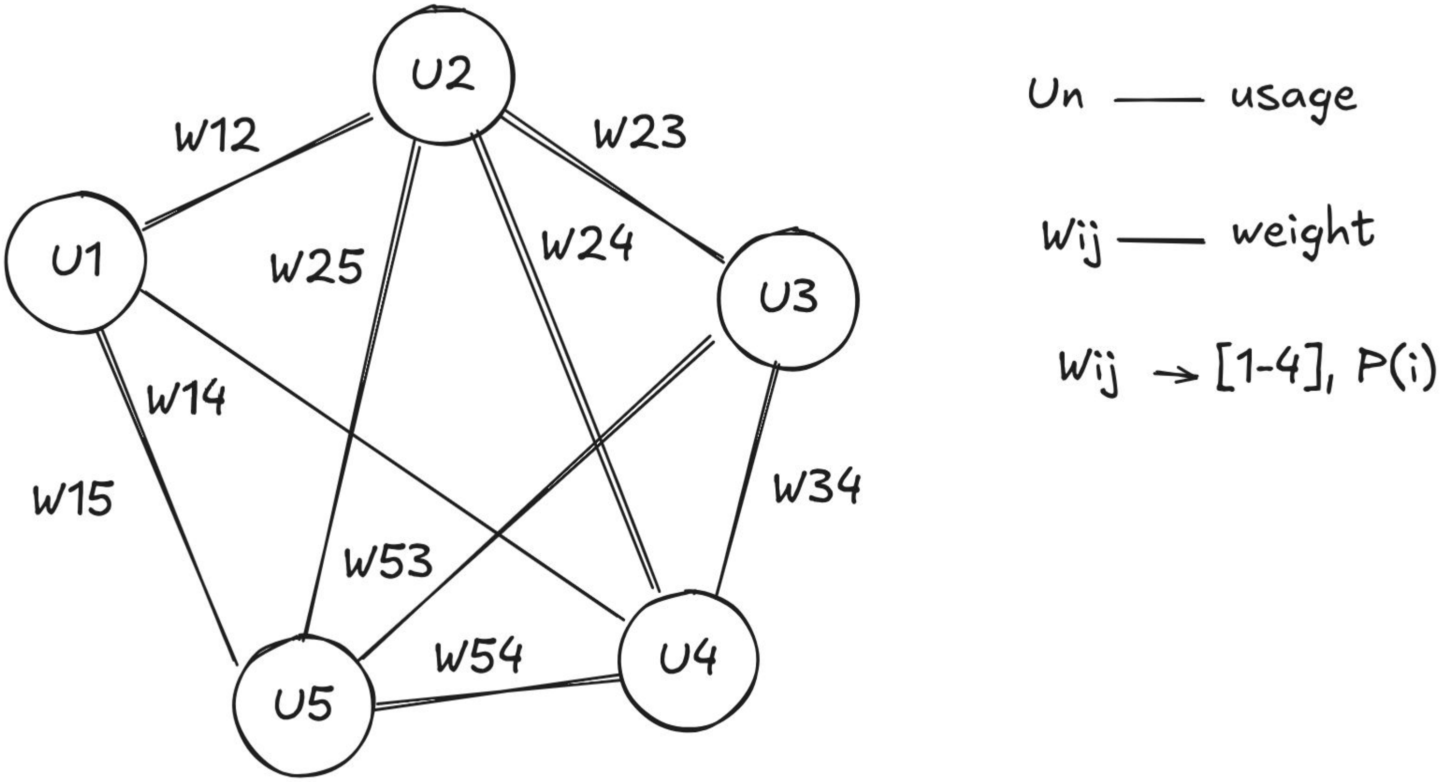


Experimental setup - WiC models



(Schlechtweg et al., 2024)

Word Usage Graph



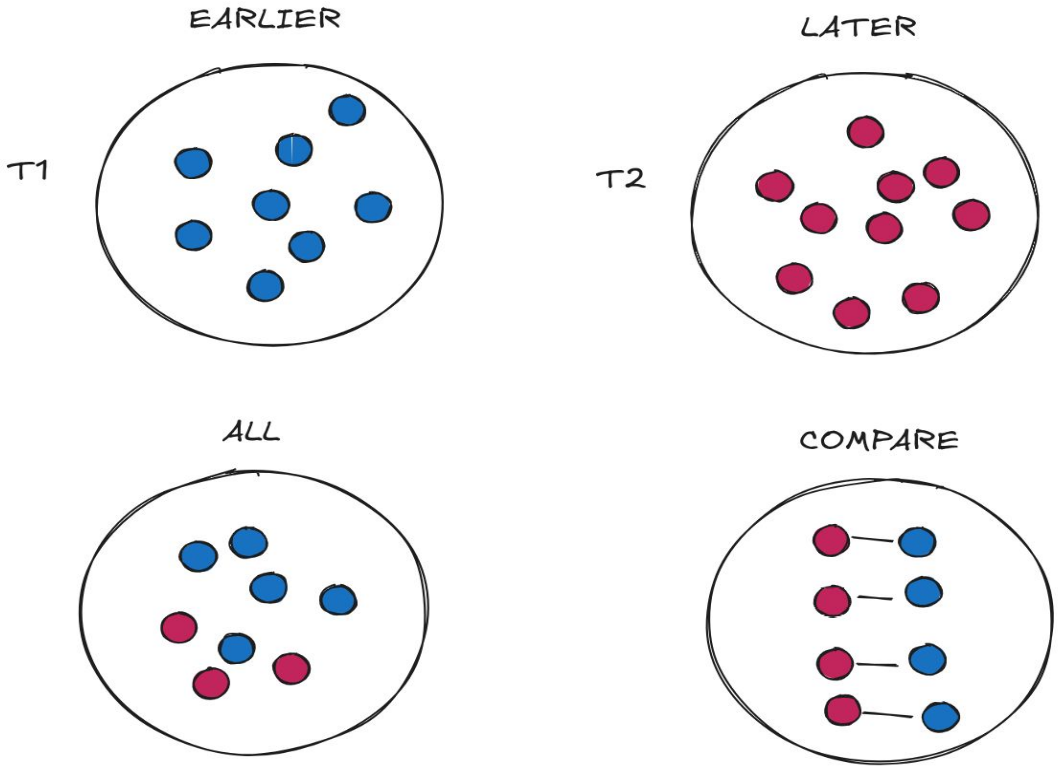
Clustering algorithms

Spectral clustering (von Luxburg, 2007)

Agglomerative clustering (Jain and Dubes,1988)

Weighted Stochastic Block Model (Peixoto, 2019)

Lexical Semantic Change Detection



Prompts

- optimized prompt from Yadav et al. (2024)
    - translate the prompt into Spanish
    - extend the prompt using samples from the DURel framework (Schlechtweg et al., 2018)
- optimized prompts further using DSPy framework (Khattab et al., 2023)

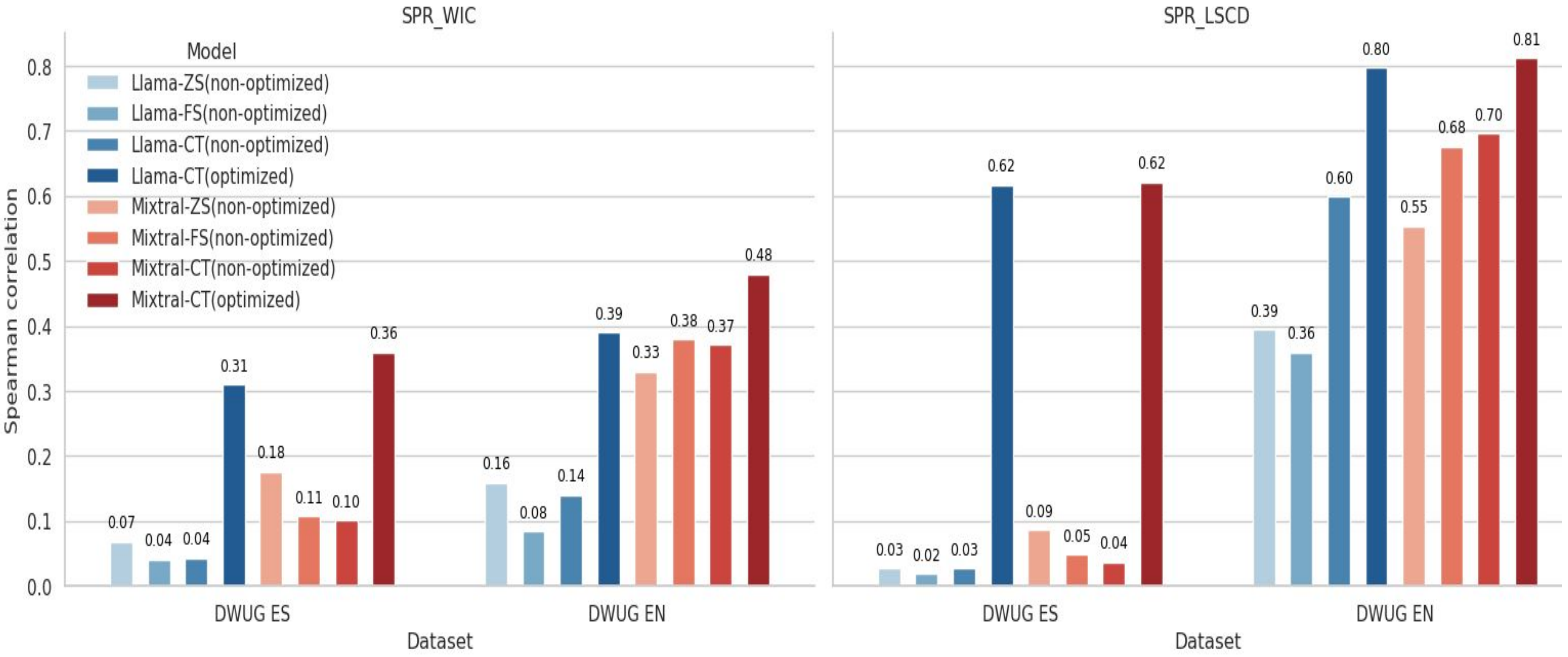
Prompt optimization

	Llama 3.1		Mixtral		Llama 3.3	
Dataset-Prompts	NOP	OP	NOP	OP	NOP	OP
DWUG ES - PrS	26.8	29.5	32.6	31.22	32.33	39.77
DWUG ES - PrE	26.5	**35.5**	33.7	**34.80**	40.88	**46.33**
DWUG EN - PrS	25.75	31.0	32.8	**40.75**	37.25	45.5
DWUG EN - PrE	28.75	**37.25**	33.25	38.75	35.75	**49.25**

RQ1

- Can automatically optimized prompts yield better results for the LSCD task than manually crafted prompts designed through prompt engineering?

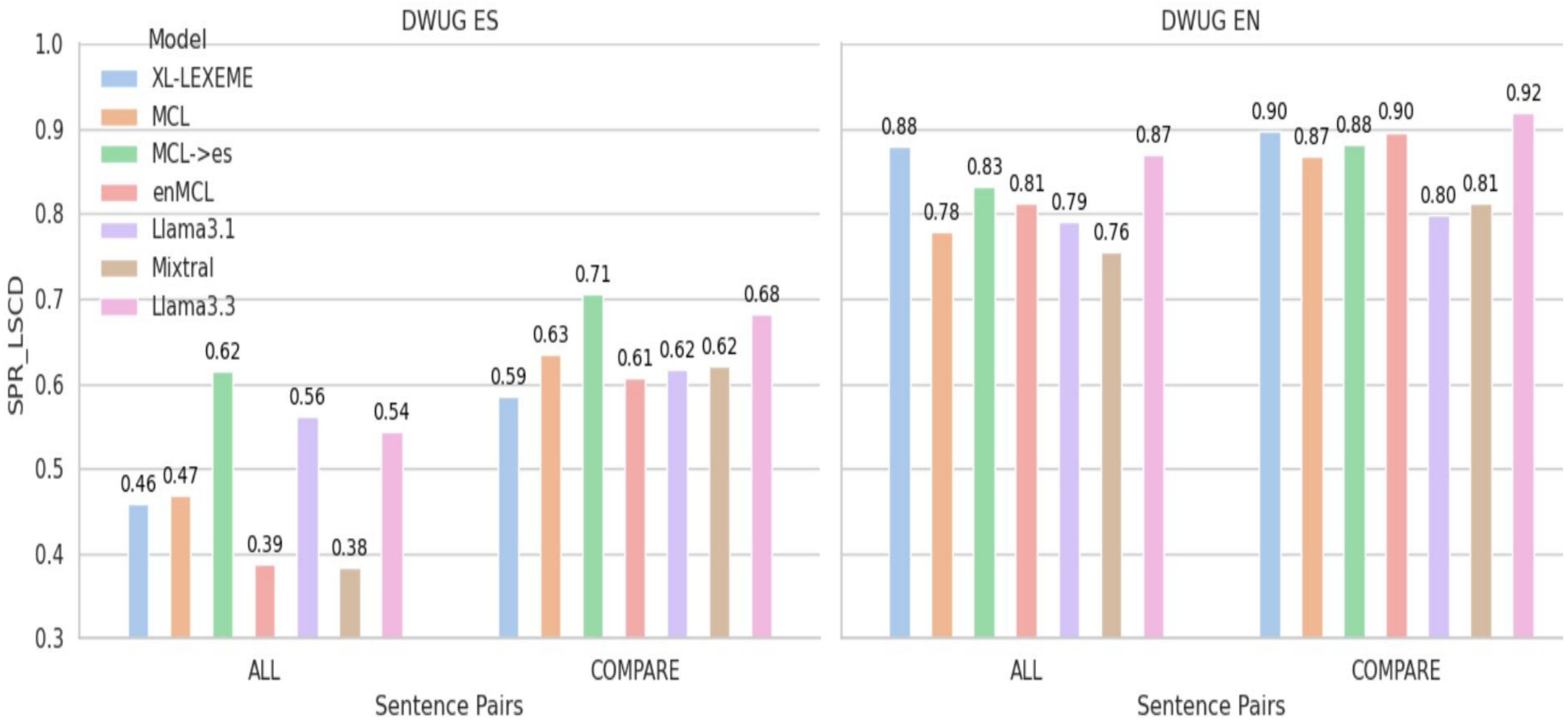
Non-optimized Prompts vs. Optimized Prompts



RQ2

- Can LLMs solve the Graded Change LSCD task well? Can these results surpass the WiC models reported as state-of-the-art?

Specialized WiC models vs LLMs in SPR_LSCD



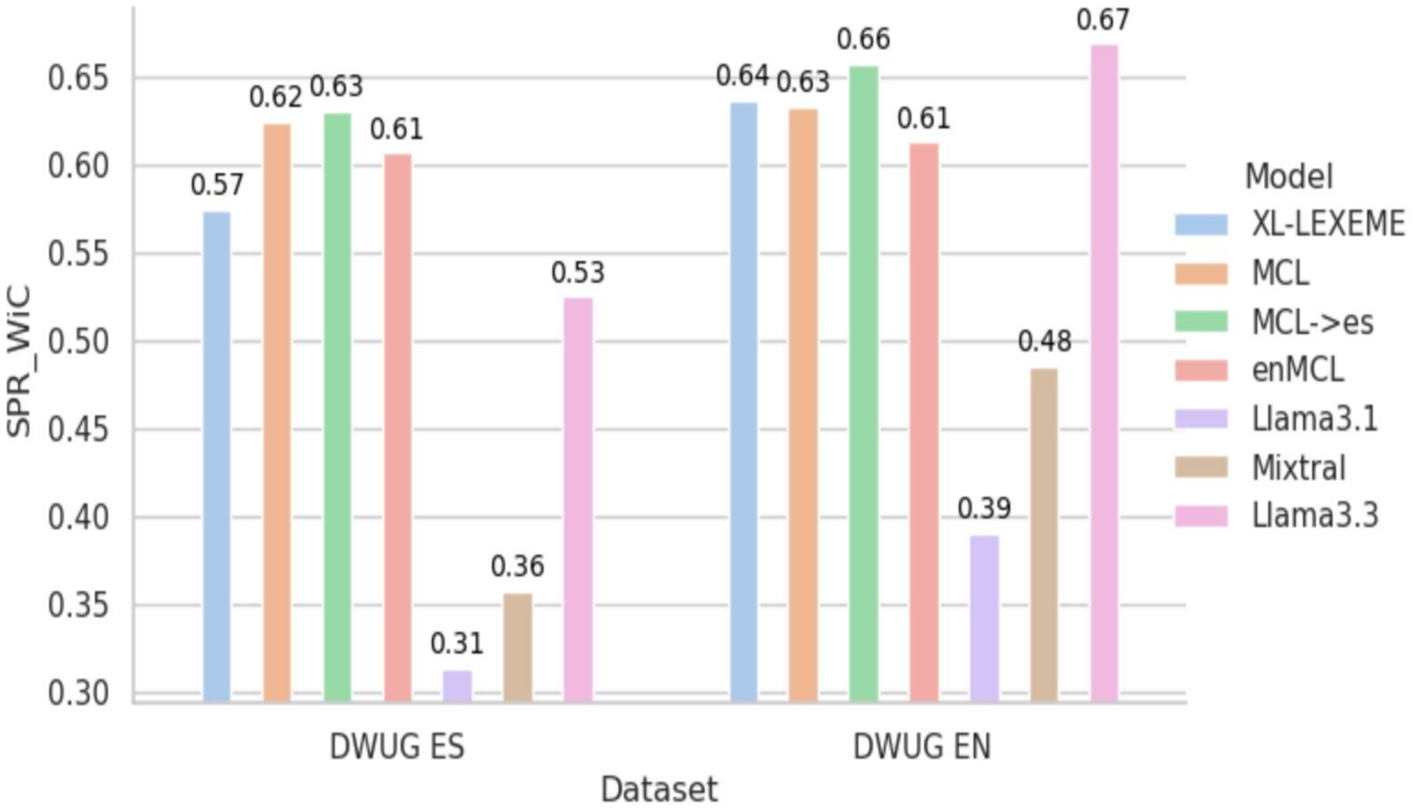
WiC + WUG + Clustering methods - cross validation

Methods	Models	Spr_LSCD (ES)	ARI (ES)	Spr_LSCD (EN)	ARI (EN)
WSBM SC AC APD	Llama 3.1	.369 ± .204 .271 ± .461 .478 ± .286 .636 ± .236	.356 ± .116 .102 ± .097 .073 ± .058 -	**.845 ± .097** .014 ± .411 .205 ± .540 .645 ± .368	.152 ± .067 -.03 ± .01 -.01 ± .03 -
WSBM SC AC APD	Mixtral	.454 ± .180 .565 ± .141 .414 ± .075 .567 ± .332	.380 ± .104 .092 ± .049 .068 ± .027 -	.776 ± .219 -.171 ± .492 -.04 ± .525 .612 ± .280	.161 ± .07 -.03 ± .01 -.003 ± .02 -
WSBM SC AC APD	Llama 3.3	.659 ± .181 .507 ± .231 .423 ± .184 .676 ± .195	**.502 ± .09** .294 ± .05 .228 ± .05 -	.729 ± .241 .302 ± .436 .195 ± .273 .752 ± .227	.183 ± .08 .124 ± .113 -.01 ± .01 -
WSBM SC AC APD	DeepMistake	**.727 ± .206** .561 ± .140 .457 ± .320 .653 ± .250	.397 ± .074 .355 ± .036 .341 ± .054 -	.730 ± .212 .520 ± .436 .433 ± .245 .638 ± .292	.231 ± .212 **.273 ± .115** .215 ± .128 -
WSBM SC AC APD	XL-LEXEME	.630 ± .377 .484 ± .215 .426 ± .255 .566 ± .354	.452 ± .095 .318 ± .043 .292 ± .087 -	.686 ± .200 .491 ± .176 .143 ± .367 .814 ± .199	.152 ± .059 .137 ± .063 .02 ± .024 -
WSBM	Random Baseline	-.199 ± .310	-.02 ± .184	-.111 ± .254	-.05 ± .147

RQ3

- Can LLMs outperform state-of-the-art LSCD models at the annotation level?

Specialized WiC models vs LLMs in SPR_WiC



Conclusions

- recent prompt optimization techniques are crucial for achieving better results on the Graded Change LSCD task, as demonstrated by the performance of Llama3.3:70B
- medium-sized LLMs such as Mixtral:8x7B and Llama3.1:8B still underperform compared to smaller and faster specialized LSCD models in both the DWUG EN and DWUG ES datasets
    - in addition to optimization techniques, the size of the model also significantly influences the results

Thanks