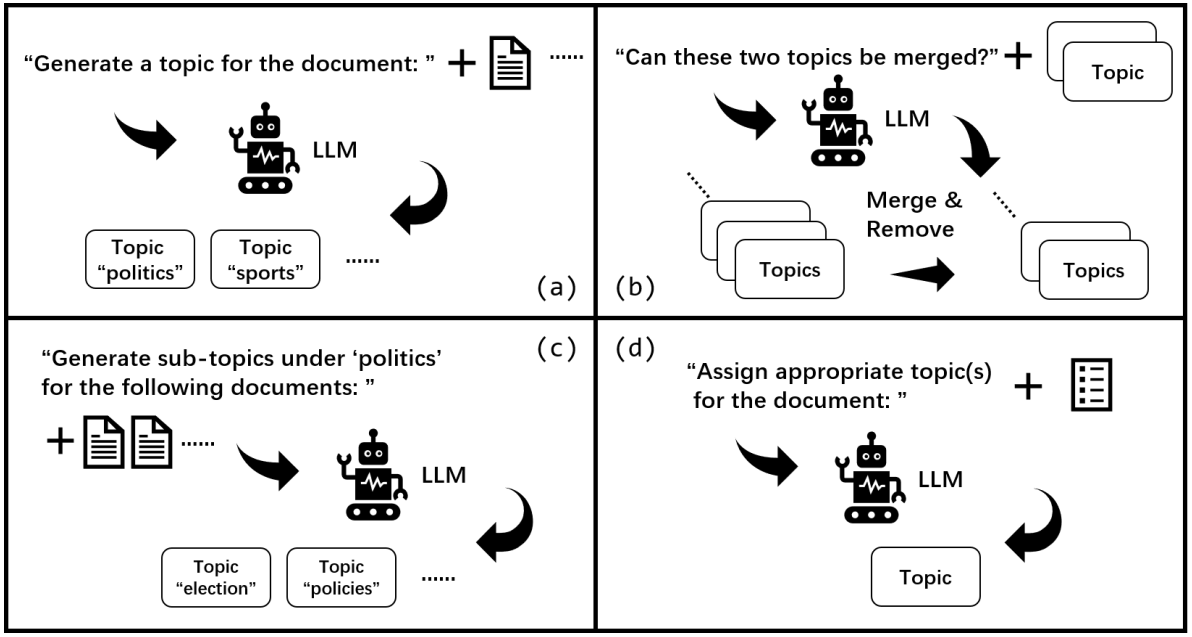
Foundation Models for Topic Modeling: A Case Study

Han ZENG, Jia-Ming SUN, Chun-Shu LI, Zhuying LI, Tong WEI

Frontiers of Computer Science, DOI: 10.1007/s11704-024-40069-7

Problems & Ideas

- Problems of conventional topic modeling approaches:
  - Failure to grasp the semantics nuances due to the BoW model
- Ideas: Utilize the profound semantic understanding ability of Large Language Models (LLMs) to enhance the efficiency and accuracy of topic modeling



Overview of out method. Key idea: Prompt the LLMs to follow human instructions in natural language on topic modeling. (a) First-level topic generation; (b) Topic refinement; (c) Second-level topic generation; (d) Topic assignment.

Main Contributions

- Contributions:
  - We used multiple LLMs for comparison, including the Qwen models deployed locally;
  - We crawled data from the Internet to build a news dataset and designed Chinese prompts to test TopicGPT in a Chinese environment;
  - We expanded the capabilities of TopicGPT and added support for third-level topic generation.

Model	$P_1 \uparrow$	ARI ↑	NMI ↑
GPT-4	76.20	65.44	83.77
Qwen-72B	72.76	54.83	81.90
Qwen-14B	55.61	19.94	73.41
LDA	52.28	13.55	52.28

Topical alignment between ground-truth labels, predicted assignments of the three models, and traditional topic modeling method LDA. Higher values indicate better performance.