richai:

I think it’s good to focus on headlines because if you have a full-text article the embedding can get weird and headlines are good compressions of the content anyways.

Headlines sometimes intentionally leave out information (clickbait,) or inflate the the significance of events. Maybe better to have an LLM create a very short summary and encode that.

Could a prompt that says “use strictly the information in the news article” help… maybe? Are there open source LLMs which are designed to be point in time… I don’t know.

LLMs should be well suited to anonymizing articles. I’d suggest you just have the LLM anonymize the article first in a separate run to strip away any information that could give away the specific date and company.