

Instruct模板

问题: {question}

请结合如下支撑信息, 回答上述问题:

支撑信息:

{retrieval}

说明: 使用所提供的支撑信息, 对给定的问题写一个全面的答复。

如果上下文不充分, 请尝试自己回答问题或者无法回答问题。

对话历史: {history}



Prompt输入

问题: 俄罗斯瓦格纳集团是一个怎样的组织?

请结合如下支撑信息, 回答上述问题:

支撑信息:

[1]瓦格纳集团 (Wagner Group) 是一个在世界各地参与xxx巴斯

[2]从普里戈任表xxxx到有关普里戈任情况的报告, 正在采取所有必要措施。北京时间6月24日 03时

说明: 使用所提供的支撑信息, 对给定的问题写一个全面的答复。

如果上下文不充分, 请尝试自己回答问题或者无法回答问题。

对话历史: {history}

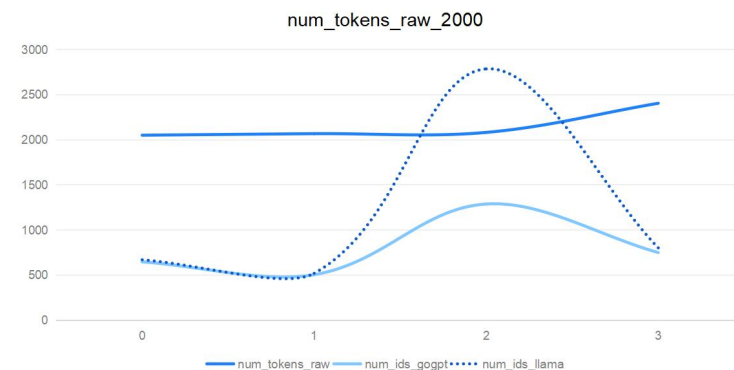
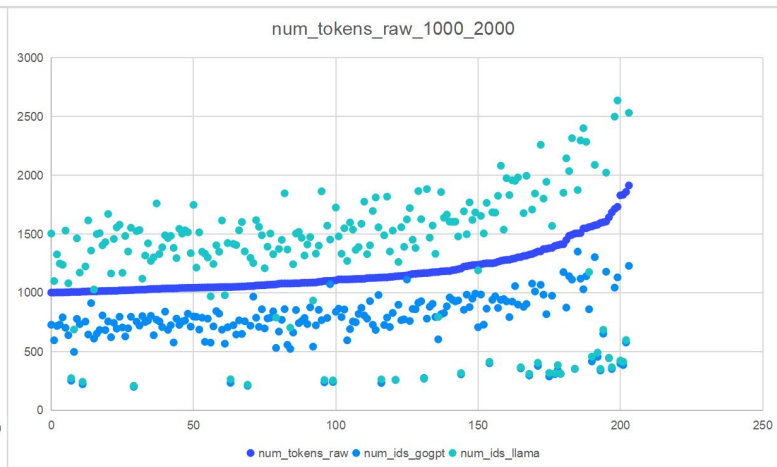
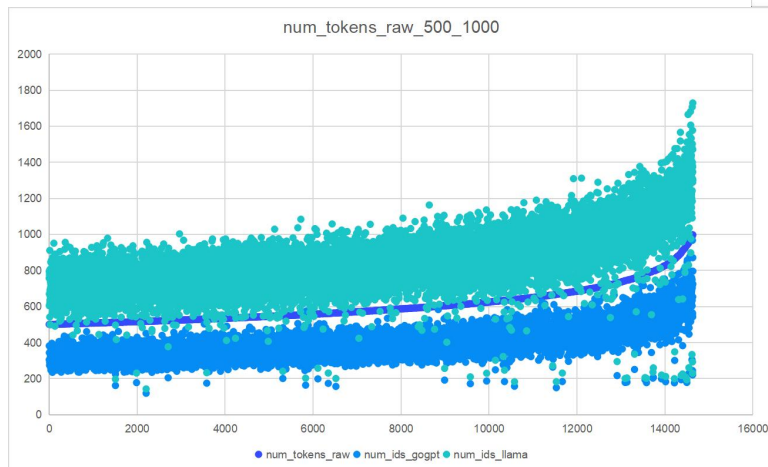
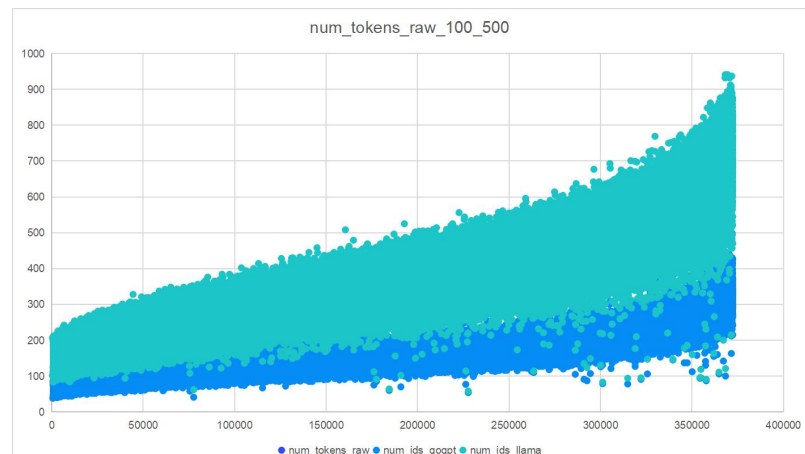
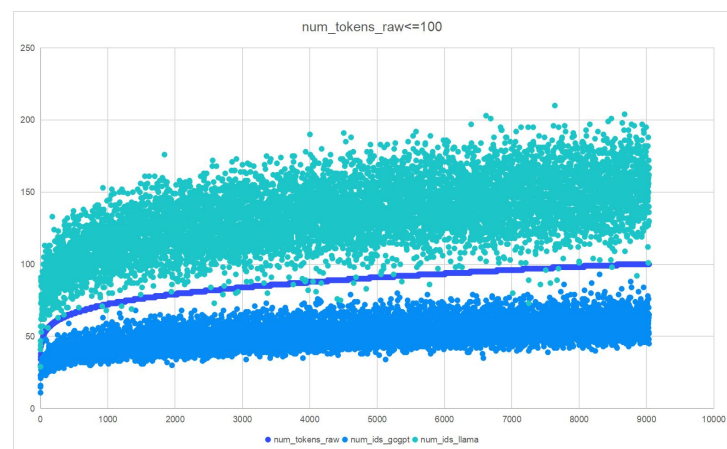
GoGPT输出



根据提供的支持信息, 可以得出以下结论:

- 瓦格纳集团是一个由多个商业网络组成的组织, 其成员主要来自于俄罗斯国防部军队、俄罗斯特种作战部队以及俄罗斯民间武装组织。其中, 一些成员是从2013年的“斯拉夫军团”中招募而来的。
- 瓦格纳集团的创始人是德米特里·乌特金, 他曾在俄罗斯国防部军队服役, 并在2014年创建了这个组织。
- 瓦格纳集团在2014年参与了“小绿人”的行动, 帮助俄罗斯军队控制克里米亚半岛。此后, 该组织还派遣约1000名雇佣军到乌克兰东部顿巴斯。
- 在2019年6月23日, 俄罗斯国防部发言人佩斯科夫发表声明, 声称俄罗斯国防部军队已经进行了打击, 并摧毁了瓦格纳雇佣军的后方营地。普里戈任作为瓦格纳雇佣军的指挥官宣称, 该营地遭受到了“巨大损失”。然而, 俄罗斯国防部随后发表声明, 指出所有关于普里戈任的信息和视频都是虚假的, 是一种信息挑衅。同时, 俄罗斯总统新闻秘书佩斯科夫也表示, 普京已收到了有关普里戈任情况的报告, 正在采取所有必要的措施。

字符个数与分词个数统计



- gogpt-tokenizer中文分词效率要明显优于llama-tokenizer
- llama-tokenizer对中文语料存在效率较低的问题，导致文本切词的序列会明显变长，超过原始文本字符个数

---num_tokens_raw:文档字符个数

---num_ids_gogpt:gogpt分词个数

---num_ids_llama:llama分词个数

---x轴代表频率，y轴代表分词个数

字符个数与分词个数统计

	num_tokens_raw	num_ids_gogpt	num_ids_llama	raw_gogpt_ratio	llama_gogpt_ratio
count	396209	396209	396209	396209	396209
mean	261.2874013	158.9009891	379.7946589	1.693523994	2.464589193
std	115.0240884	80.63505636	166.3046338	0.205930733	0.327637156
min	23	11	29	0.896296296	0.974545455
30%	194	112	282	1.573770492	2.296610169
50%	242	142	351	1.695652174	2.46031746
90%	401	257	589	1.952830189	2.879032258
max	2405	1350	2786	5.251396648	4.371428571

- 90%文档的字符个数约在400-500字左右;
- 90%情况下gogpt对原始文本字符压缩率为1.95，约等于2倍;
- gogpt模型上下文窗口大小为2048，可接受常规文档的个数约为9-12个，约4000字;
- gogpt分词器相比llama原始分词器，对中文平均压缩率约为0.41(1/2.46)，训练和推理效率显著提升

字符个数与分词个数统计

500字直观上是怎样的？

成长，因有梦而精彩

她站在灿烂的阳光下，望着深远的天空，突然想起了一个字：梦。她的成长，正是因为有梦，才如此精彩。小时候，她很喜欢在梦中畅游。她想象自己是英雄，又想象自己是受小孩喜爱的卡通人物。长大些，她渐渐地明白，梦，并非想象，而是一种能催人奋发向上的力量，一个能让人为之努力的目标。不知在何时，她找到了属于她自己的梦——作家。她明白，梦总是遥远的，却不是虚无飘渺的；她明白，梦总是神圣的，却不是触不可及的。她为梦付出了很多，她的成长也因此而精彩。因为有梦，她选择坚持。小学时，课余时间很丰富。别的同学都在疯狂地玩耍，她也参与其中。不同的是，她总在默默地观察着一切，从中发现各种各样的“美”。大自然的美，亲情的美，友谊的美……都令她为之感动。她欣喜地将这一切记录下来，从不感到疲倦，也从不感到厌烦。她就这样坚持着，因为她有梦。因为有梦，她选择坚强。上初中后，黑夜来了。一个又一个困难从天而降，她一时被困难击得站不起来。她感到力不从心，学习科目的突然增多，让她毫无防备，顿时乱了套；作业的增多、任务的加重让她难以应付。她甚至不想再站起来，站到队伍的最前列。可是，她心中的梦在爆发，在跃动，在奔腾，她强忍疼痛，站了起来，并加快速度向无尽的前方走去。在日落时，她选择了坚强，因为她有梦。因为有梦，她选择永不言弃，她选择坚持不懈，她选择奔跑向前。她，也就是我，又一次仰望着深远的天空，面对着初升的红日，大声地喊出：“成长因有梦而精彩。”