

# 前沿语言大模型性能对比测试

彭海琅 ZY2203322

plunck@buaa.edu.cn

## 摘要

人工智能如何发展出像人类具备逻辑、意识和推理的认知能力，是人工智能研究一直探索的方向。目前来看，通过大规模数据训练超大参数量的巨量模型，被认为是非常有希望实现通用人工智能的一个重要方向。AI 大模型是指一个庞大复杂的神经网络，需要通过存储更多的参数来增加模型的深度和宽度，从而提高模型的表现能力，参数从百亿起步，对大量数据进行训练并产生高质量的预测结果。本文从文学创作、知识问答、新闻撰写、中译英、数学推理五个下游任务的性能来对比三个优秀的大语言模型——ChatGPT、NewBing、GPT-4，总的来说，GPT-4 的性能最优秀。

## 导论

### 一、AI 大模型

AI 大模型是指一个庞大复杂的神经网络，需要通过存储更多的参数来增加模型的深度和宽度，从而提高模型的表现能力，参数从百亿起步，对大量数据进行训练并产生高质量的预测结果。AI 大模型能够应用到很多领域，包括自然语言处理、计算机视觉、语音识别等。在自然语言处理领域，AI 大模型可以用于自动问答系统、语音助手等；在计算机视觉领域，AI 大模型可以用于图像识别、人脸识别等；在语音识别领域，AI 大模型可以用于语音识别、自动翻译

等。此外，AI 大模型还可以应用于金融、医疗、教育等行业，帮助人们提高工作效率和生活质量。

在 2019 年发布的 GPT-2 被认为是第一个大型语言模型 (Large Language Model, LLM)，拥有 15 亿个参数，培训成本估计为 5 万美元。三年后，PaLM 推出，拥有 5400 亿个参数，培训成本估计为 800 万美元。各个领域的大型语言和多模态模型都变得越来越大和昂贵。这意味着人工智能系统的能力正在不断提高，需要更困难的测试才能挑战它们。中国工程院院士王恩东表示：“人工智能如何发展出像人类具备逻辑、意识和推理的认知能力，是人工智能研究一直探索的方向。目前来看，通过大规模数据训练超大参数量的巨量模型，被认为是非常有希望实现通用人工智能的一个重要方向。”随着巨量模型的兴起，巨量化已成为未来人工智能发展非常重要的一个趋势。

## 二、下游任务

由于 NLP 领域面对的问题太过庞杂，因此前面很多年 NLP 领域任务都被拆分的非常细，包含但不限于以下几类：

- 命名实体识别 (Named Entity Recognition, NER)：比如对于输入语句「擎天柱回到赛博坦」得到输出「B-PER, I-PER, E-PER, O, O, B-LOC, I-LOC, E-LOC」，其中 B、I、E 分别表示开始、中间、结束，PER、LOC 分别表示人物、地点，O 表示其他无关。
- 文本蕴含 (Text Entailment)：比如对于文本 T「我在杭州」和如下三个假设 H1「我在浙江」、H2「我在上海」、H3「我是杭州人」之间的蕴含关系就是 Positive、Negative、Neutral，其实是个三分类问题。
- 常识推理 (Common Sense Reasoning)：比如一个测试 LM 是否具备常

识推理的例子，在句子 A「奖杯无法放进到箱子里，因为它太了」中的「它」指的谁？在句子 B「奖杯无法放进到箱子里，因为它大了」中「它」指的谁？这样的推理依赖于常识。

- 问答 (Question Answering)。
- 词性标注 (POS Tagging)。
- 情感分析 (Sentiment Analysis, SA)。
- 自然语言推理 (Natural Language Inference, NLI)。
- 总结摘要 (Summarization)。
- .....

图 1 整理了面向 NLP 领域不同下游任务的部分数据集。

### 三、提示工程

对于 ChatGPT 这样的大型语言模型来说，提示 (Prompt) 是非常重要的，因为 Prompt 直接影响模型的生成结果和质量，Prompt 可以帮助模型更好地理解你的意图和需要，并生成更加准确和有针对性的回复。此外，Prompt 还可以帮助模型更好地处理多义词、歧义和上下文依赖性问题，提高模型的准确性和自然度。设计最佳提示以指导模型完成任务的这种方法被称为提示工程 (Prompt engineering)。

早在 2020 年，OpenAI 就在论文 Language Models are Few-Shot Learners 中提出了如何使用 prompt learning 提升大模型的推理能力。论文中提出了 Zero-shot、One-shot、Few-shot 三种不同的 prompt 方法，如图 2 所示。

Few-Shot (FS) 是指模型在推理时给予少量样本，但不允许进行权重更新。对于一个典型数据集，Few-shot 有上下文和样例（例如英语句子和它的法

Type	Task	Datasets	Class	Length	Number	Related Papers
Classification	Sentiment Analysis	MR	2	20	10662	[403, 404, 405, 406, 407]
		SST-1	5	18	11,855	[408, 403, 409, 410, 411]
		SST-2	2	19	9,613	[408, 403, 412, 413, 13]
		MPQA	2	3	10,606	[414, 403, 415]
		IMDB	2	294	50,000	[416, 417, 412, 413, 418, 14]
	News Classification	20NG	20	221	18,846	[419, 420, 421, 406, 422, 279]
		AG News	4	45/7	127,600	[423, 424, 425, 405, 14]
		R8	8	66	7,674	[406, 422, 426]
		R52	52	70	9,100	[406, 422, 426]
	Topic Labeling	DBPedia	14	55	630,000	[423, 424, 418, 427]
		Ohsumed	23	136	7,400	[406, 422, 426]
		YahooA	10	112	1,460,000	[423, 428]
	Natural Language Inference	SNLI	3	-	570,152	[429, 430, 55, 431, 13, 275]
		MNLI	3	-	433,000	[432, 13, 14, 55, 36]
		QNLI	2	-	115,667	[13, 14, 36]
		WNLI	2	-	852	[431, 36]
		RTE	2	-	5,768	[36]
		SICK	3	-	10,000	[433]
		MSRP	2	-	5,801	[434]
	Named Entity Recognition	CoNLL 2003	4	-	2,302	[275, 13, 435, 436, 437, 438]
		OntoNotes 4.0	18	-	-	[439, 440]
		OntoNotes 5.0	18	-	2,945,000	[13, 435, 436, 438]
		MSRA	3	-	-	[439, 13, 440, 438]
		ACE 2004	7	-	443	[441, 442, 443, 444, 438]
		ACE 2005	7	-	437	[441, 442, 443, 445, 438]
		KBP2017	-	-	-	[445, 438]
	Question Answering	QQP	2	-	799,266	[13, 36]
		MRPC	2	-	-	[36]
		SQuAD	-	5,000	5,570	[275, 55, 36]
		RACE	5	-	100,000	[446, 14, 431, 36]
		TREC	6	10	6,400	[404, 412, 425, 279, 405, 427]
		WikiQA	-	873	243	[447, 448]
		DSTC 4	89	-	30,000	[449, 450]
	Dialog Act Classification	MRDA	5	-	62,000	[451, 449]
		SwDA	43	-	1,022,000	[449, 452, 453]
Generation	Text Summarization	NYT	-	-	109,910	[454, 455]
		CNN	-	760	92,579	[456, 457, 458, 459, 460]
		Dailymail	-	653	219,506	[461, 457, 454, 462, 459]
		Gigaword	-	-	3,991,000	[463, 457]
	Machine Translation	WMT14	-	-	-	[464, 465]
		WMT16	-	-	-	[466, 465]
		WMT17	-	-	-	[467, 468, 466, 464, 469]
		WMT18	-	-	-	[467, 466, 468]
	Dialogue	DSTC2	-	-	3,000	[470]
		MWOZ	35	15.03	10,438	[470, 471, 472]
		GSIM	-	-	3,008	[470]
		OOS	151	-	23,700	[470]

图 1 不同下游任务的数据集整理

语翻译)。Few-shot 的工作方式是提供  $K$  个样本，然后期望模型生成对应的结果。通常将  $K$  设置在 10 到 100 的范围内，因为这是可以适应模型上下文窗口的示例数量 ( $nctx = 2048$ )。Few-shot 的主要优点是大幅度降低了对特定任务数据的需求，并减少了从微调数据集中学习过度狭窄分布。主要缺点是该方法的结果迄今为止远不如最先进的微调模型。此外，仍需要一小部分特定任务的数据。One-Shot (1S) 与 Few-Shot 类似，只允许一个样本（除了任务的自然

语言描述外)。将 One-Shot 与 Few-Shot、Zero-Shot 区分开的原因是它最接近某些任务与人类沟通的方式。相比之下，如果没有示例，有时很难传达任务的内容或格式。Zero-Shot (0S) 和 One-shot 类似，但不允许提供样本，只给出描述任务的自然语言指令。该方法提供了最大的方便性、稳健性以及避免虚假相关的可能性，但也是最具挑战性的设置。在某些情况下，即使是人类，在没有例子的情况下，也可能难以理解任务的格式。例如，如果要求某人“制作一张关于 200 米冲刺世界纪录的表格”，这个请求可能是模棱两可的，因为可能不清楚表格应该具有什么格式或包含什么内容。然而，至少在某些情况下，Zero-shot 是最接近人类执行任务的方法，例如图 2 中的翻译示例，人类可能仅凭文本指令就知道该做什么。

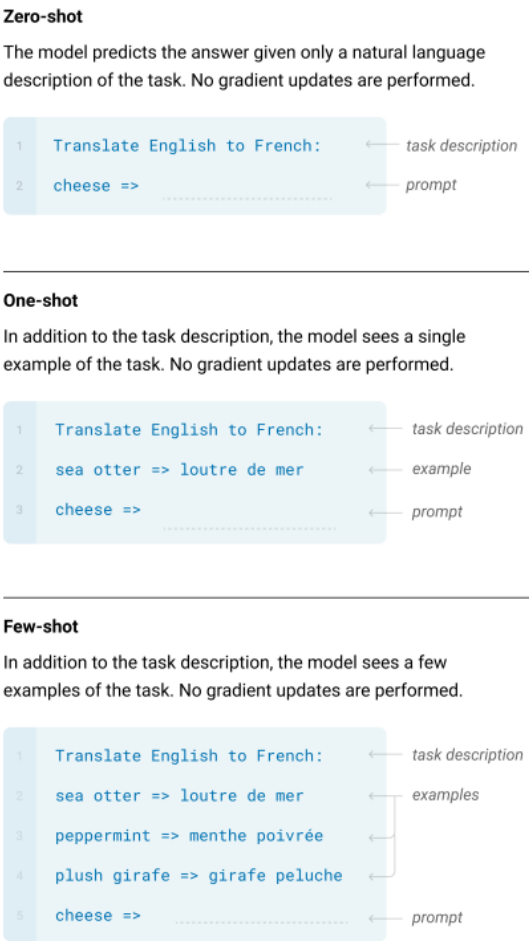


图 2 Zero-shot, One-shot and Few-shot

一般来说，我们在撰写提示时可以做到以下几点：

- 具体性：请具体说明您想要模型执行的指令和任务。提示越描述详细，结果就会越好。当您需要的是特定的生成结果或风格时，这一点尤为重要。没有特定的令牌或关键字能够提供更好的结果。拥有良好的格式和详细的提示更为重要。在提示中提供例子非常有助于在特定格式下获得所需的输出。
- 避免不精确：给出了有关详细说明和改进格式的提示之后，很容易陷入关于提示过于聪明并可能创建不准确的描述的陷阱。通常更好的方法是具体而直接。这里的类比与有效的沟通非常相似——越直接，信息传递就越有效。
- 指出应该做什么：设计提示时的另一个常见提示是避免说出不要做什么，而要说出的做法。这样可以更具体，并侧重于细节，这些细节可以引导模型产生良好的响应。

## 实验内容

### 一、实验方法

在 Chat-GPT、GPT-4、NewBing 三个大模型上分别测试文学创作、知识问答、新闻撰写、中译英、数学推理五个下游任务的能力，对三个大模型的性能进行比较评判。

### 二、实现步骤

#### 1、文学创作测试。

Prompt：我想让你扮演一个讲故事的人。你会想出引人入胜、富有想象

力、引人入胜的有趣故事。它可以是童话、教育故事或任何其他类型的故事，有可能吸引人们的注意力和想象力。根据目标受众的不同，你可以选择讲故事的特定主题或主题，例如，如果是孩子，那么你可以谈论动物；如果是成年人，那么基于历史的故事可能会更好地吸引他们。我的第一个要求是“我需要你对孩子们讲一个关于毅力的有趣故事。”

Chat-GPT 的回复：



好的，我将为您讲一个有趣的故事，它关于一只叫做毅力的小狗。

毅力是一只可爱的小狗，它住在一座小村庄里。毅力非常聪明，而且非常喜欢玩耍。每天早晨，它都会去找其他小狗一起玩耍，但它总是被嘲笑和冷落。

然而，毅力并不放弃。它知道它需要找到一种方法来赢得其他小狗的尊重和友谊。于是，他开始观察其他小狗怎么玩，并在他们身上找到了一种共同的特点：他们都非常有毅力！

于是，毅力开始学习如何利用自己的特质。它开始每天早上早早地起床，去练习跑步、跳跃和翻滚等技能。虽然它经常会摔倒，但它永远不会放弃，它会一直练习，直到它掌握这些技能。

其他小狗看到了毅力的努力和坚持，开始感到惊叹和敬佩。他们开始和毅力一起玩耍，并尊重他的实力。他们发现，毅力不仅仅是一只可爱的小狗，他还是一位真正的毅力大师！

通过毅力的努力和坚持，他终于赢得了其他小狗的尊重和友谊。他也成为了他们中的一员，他们一起玩耍、探险和成长。

这个故事告诉我们，只需要一点毅力和坚持，我们就可以克服任何困难，实现我们的目标。无论我们想要达成什么目标，只要我们相信自己的能力，并坚持不懈地努力，我们就可以成功！

NewBing 的回复：

好的，我可以讲一个关于毅力的故事。这个故事叫做“乌龟和兔子赛跑”。在这个故事中，兔子以为他很快，所以他不认为自己需要努力。相反，乌龟知道他不是最快的，但他知道他可以坚持下去。在比赛中，兔子开始跑得很快，但他很快就累了，所以他决定休息一下。乌龟一直坚持不懈地前进，最终赢得了比赛。这个故事告诉我们，即使你不是最聪明或最快的人，只要你坚持不懈地努力工作，你就可以获得成功。

GPT-4 的回复：

从前，在一个遥远的森林里，有一群动物住在一起。他们每天都在森林里快乐地玩耍。在这个大家庭里，有一只名叫小毅的刺猬。小毅非常友好，总是愿意帮助别人。但是，他有一个问题：他总是很容易放弃。

有一天，森林里的动物们决定举行一场比赛，看谁能爬到附近一座高山的顶端。所有的动物都参加了这个比赛，包括小毅。比赛开始了，动物们争先恐后地往山顶爬去。

在比赛的过程中，有些动物很快就感到疲惫。大象觉得自己太重了，很快就放弃了。长颈鹿发现它的脖子太长，无法稳定地攀爬。猴子虽然动作敏捷，但是却被山上的悬崖吓到了。很多动物都因为各种原因放弃了比赛。

小毅在爬山的过程中，也遇到了很多困难。他的脚短，爬山时速度很慢。山上的岩石又滑又陡，让他很难攀爬。但是，小毅想起了他的朋友们，他们告诉他毅力很重要，要坚持到底。于是，小毅咬紧牙关，继续向前爬。

就这样，小毅一步一步地往上爬，不管遇到多大的困难，他都没有放弃。渐渐地，小毅发现，他已经超过了许多其他动物。最后，只剩下小毅和一只名叫飞翔的雄鹰。

飞翔是一只非常自信的雄鹰，他认为自己一定能轻松赢得比赛。但是，他看到小毅毫不气馁地爬上来，感到非常惊讶。飞翔开始担心，他怀疑自己是否真的能赢得比赛。

最后，小毅和飞翔同时到达了山顶。其他动物都为它们欢呼，因为这是一场非常激动人心的比赛。小毅终于明白了毅力的重要性，他知道只要不放弃，就一定能实现自己的目标。

从那以后，小毅变得更加坚定，他的毅力感染了所有的动物。他们都学会了不轻易放弃，勇往直前。而小毅也成为了森林里最受尊敬的动物之一。这个故事告诉我们，只要有毅力，就没有什么困难是无法克服的。所以，孩子们，无论遇到什么挑战，都要像小毅一样，坚持不懈，勇往直前。这样，我们才能实现自己的梦想。

## 2、知识问答测试。

Prompt：我希望你能成为一名医生，为疾病或疾病想出创造性的治疗方法。你应该能够推荐传统药物、草药和其他天然替代品。在提供建议时，您还需要考虑患者的年龄、生活方式和病史。我的第一个建议是“为患有螨虫过敏性鼻炎的年轻患者制定一个专注于整体治疗方法的治疗计划”。

Chat-GPT 的回复：



作为一名医生，我非常愿意为您提供建议。对于螨虫过敏性鼻炎的年轻患者，我建议制定一个专注于整体治疗方法的治疗计划，以改善他们的健康状况和生活质量。

在这个治疗计划中，我建议采取以下措施：

1. 药物治疗：药物治疗是治疗螨虫过敏性鼻炎的主要方法。我建议采用抗组胺药和类固醇鼻喷雾剂来减轻症状和缓解炎症。同时，我建议患者在日常生活中采取预防措施，如避免过敏原、保持室内空气流通和定期清洁家居等。
2. 免疫疗法：免疫疗法也是一种有效的治疗方法，它通过慢慢注射过敏原来减少人体对过敏原的敏感性。这种治疗方法通常在医院进行，需要较长时间的进行。
3. 中草药治疗：中草药在治疗螨虫过敏性鼻炎方面也具有一定的疗效。例如，菊花、金银花和蒲公英等草药都具有清热解毒、祛风止痒的功效，可以帮助缓解症状。
4. 心理治疗：螨虫过敏性鼻炎常常伴随着焦虑和抑郁等心理问题。因此，我建议患者接受适当的心理治疗，以帮助他们处理鼻炎带来的不适和情绪问题。

需要注意的是，治疗计划应该根据患者的具体情况进行定制。因此，我建议您与患者的医生进行密切合作，以确保治疗方案最适合患者。同时，患者也应该积极配合治疗，坚持日常生活中的预防措施，以达到最佳治疗效果。



NewBing 的回复：

你好，对于螨虫过敏性鼻炎的治疗方法，建议患者朋友尽量减少居室内的螨虫浓度，勤清洗被褥，保持通风，用防螨床垫，通过这种方法减少居室内螨虫的浓度。药物治疗包括抗组胺药、鼻用激素、白三烯受体拮抗剂等，根据临床症状，在医生建议下选择性使用 <sup>1</sup>。

如果您想了解更多关于尘螨过敏性鼻炎的治疗方法，建议您咨询专业医生以获取更详细的信息。 <sup>2</sup>

希望这可以帮助你。请问还有什么其他问题吗？

GPT-4 的回复：

作为一名医生，针对患有螨虫过敏性鼻炎的年轻患者，我会制定一个整体治疗方案。这个方案将包括药物治疗、生活方式调整和使用天然替代品。以下是我的建议：

药物治疗：

口服抗过敏药物：如非处方的抗组胺药（如洛拉塔丁、西替利嗪等），可以减轻症状。

鼻喷剂：鼻腔类固醇喷剂（如氟替卡松、布地奈德等）可以减轻鼻炎症状。

生活方式调整：

定期清洁家居环境，尤其是床上用品、地毯和窗帘，以减少螨虫的滋生。

使用除螨剂和防螨床垫、枕头等产品。

保持室内湿度在40-50%之间，以减少螨虫生长的可能性。

避免抽烟和二手烟暴露，这会加重过敏症状。

增强锻炼，提高身体免疫力。

天然替代品：

鼻腔冲洗：使用生理盐水冲洗鼻腔，有助于清除过敏原，并缓解鼻炎症状。

花粉：蜂花粉是一种天然的抗组胺剂，可以减轻过敏症状。但请在医生指导下使用，因为花粉可能会导致某些人过敏反应。

薰衣草精油：薰衣草精油具有抗炎和抗过敏特性，可以通过熏香或局部涂抹的方式使用。


在治疗过程中，患者需要定期复查，以评估治疗效果并进行相应调整。请注意，本建议仅供参考，具体治疗方案请遵循专业医生的建议。

3、新闻撰写测试。

Prompt: 你是 X 市的一名记者，请你写一篇 1000 字左右的新闻，主题是：

全市政法系统会议召开。

Chat-GPT 的回复：

 X 市政法系统会议隆重召开

近日，X 市政法系统会议在市政府会议室隆重召开。会议旨在深入学习贯彻中央和省政法工作会议以及全国市域社会治理现代化工作会议精神，总结去年工作，分析当前形势，部署今年任务。

市委书记、市长、市政法委书记等领导出席会议，全市政法系统的干部职工、派出所所长、法庭庭长等共计 1000 余人参加了会议。

会议由市委副书记、政法委书记主持。他首先传达了中央和省政法工作会议以及全国市域社会治理现代化工作会议精神，并就 X 市贯彻落实工作提出了具体要求。

市委书记在会上发表了重要讲话。他强调，政法工作是国家安全工作的重要组成部分，是维护国家安全和社会稳定的重要力量。过去一年，X 市的政法工作取得了显著成绩，特别是在疫情防控方面，全市政法系统充分发挥了职能作用，为打赢疫情防控阻击战作出了重要贡献。

市委书记指出，今年是我国发展进程中极为重要的一年，也是 X 市全面建设社会主义现代化城市的开局之年。全市政法系统要认真贯彻落实中央和省委的决策部署，紧紧围绕“建设现代化城市、创造高品质生活”的目标，努力推动政法工作迈上新台阶。

市委副书记、政法委书记在主持会议时强调，全市政法系统要深刻领会市委书记的讲话精神，进一步加强自身建设，提升能力水平，为 X 市全面建设社会主义现代化城市提供坚强的政法保障。

会议还表彰了 X 市政法系统的优秀个人和先进集体，并举行了揭牌仪式，为新成立的 X 市平安建设办公室、X 市反邪教斗争委员会、X 市网络安全和信息化委员会办公室等机构揭牌。

据悉，X 市政法系统将在新的一年，紧紧围绕“建设现代化城市、创造高品质生活”的目标，始终坚持以人民为中心，全力维护国家安全和社会稳定，推动市域社会治理现代化，努力建设和谐、稳定、有序的法治 X 市。

## NewBing 的回复：

好的，以下是我为您写的新闻：

X市政法系统会议于6月7日上午在市政府会议室召开。市委政法委书记李明主持会议。市政法系统各级领导、各部门负责人、各区县政法委书记、各街道办事处主任等参加了会议。

会议主要包括：全市政法系统贯彻落实党的十九届六中全会精神，总结2022年度工作，安排部署2023年度工作。李明书记在会上强调，全市政法系统要深入学习贯彻习近平总书记关于政法工作的重要指示精神，坚持以人民为中心的发展思想，切实维护首都政治安全和社会稳定。

此外，李明书记还就全市政法系统“不忘初心、牢记使命”主题教育工作作出了重要讲话。他强调，全市政法系统要深入学习贯彻习近平总书记关于“不忘初心、牢记使命”重要论述，以主题教育为契机，推动全市政法工作高质量发展。

以上是我为您写的新闻，希望对您有所帮助。如果您有其他问题，请随时问我。

了解详细信息: [1. zhuanlan.zhihu.com](http://1.zhuanlan.zhihu.com) [2. wiki.mbalib.com](http://2.wiki.mbalib.com) [3. zhuanlan.zhihu.com](http://3.zhuanlan.zhihu.com)

5 共 5 ●

## GPT-4 的回复：

新闻标题：全市政法系统会议召开，努力维护社会安全稳定  
X市，月日讯（记者 张三）今天上午，全市政法系统会议在市政府会议中心隆重召开。市委常委、政法委书记李四主持会议并讲话，市委副书记、市长王五出席会议并讲话。会议旨在深入贯彻落实党的十九大精神，全面总结过去一年政法工作，部署2022年政法工作重点任务，努力维护社会安全稳定。  
会议首先传达了党的十九大精神，以及中央、省委有关文件精神。会议指出，党的十九大明确了新时代我国社会主要矛盾，为新时代政法工作提供了基本遵循。全市政法系统要深入学习贯彻习近平新时代中国特色社会主义思想，紧紧围绕全市经济社会发展大局，坚决维护国家政治安全，全面加强社会治理，确保人民安居乐业、社会安定有序。  
市委常委、政法委书记李四在讲话中指出，过去一年，全市政法系统紧紧围绕中心，服务大局，坚决维护国家政治安全，全面加强社会治理，确保人民安居乐业、社会安定有序。一是深入推进扫黑除恶专项斗争，严厉打击涉黑涉恶犯罪，坚决铲除黑恶势力滋生土壤。二是全面加强社会治理，深入推进基层治理体系建设，切实解决群众反映强烈的突出问题。三是严密防范化解重大风险，强化金融、土地、环保等领域监管，确保经济社会大局稳定。四是加强政法队伍建设，严明政法纪律，提高政法工作科学化、专业化、精细化水平。  
市委副书记、市长王五在会上强调，新的一年，全市政法系统要深入贯彻落实党的十九大精神，紧紧围绕新时代新任务新要求，坚决维护国家政治安全，全面加强社会治理，确保人民安居乐业、社会安定有序。一是深化扫黑除恶专项斗争，严厉打击涉黑涉恶犯罪，坚决铲除黑恶势力滋生土壤。二是全面加强社会治理，深入推进基层治理体系建设，切实解决群众反映强烈的突出问题。三是严密防范化解重大风险，强化金融、土地、环保等领域监管，确保经济社会大局稳定。四是加强政法队伍建设，严明政法纪律，提高政法工作科学化、专业化、精细化水平。  
会议最后，李四对全市政法系统2022年工作提出明确要求：一是深入贯彻落实党的十九大精神，坚定不移全面从严治党，加强政法队伍建设。二是紧紧围绕全市经济社会发展大局，全面加强社会治理，确保人民安居乐业、社会安定有序。三是深入推进扫黑除恶专项斗争，严厉打击涉黑涉恶犯罪，坚决铲除黑恶势力滋生土壤。四是严密防范化解重大风险，强化金融、土地、环保等领域监管，确保经济社会大局稳定。  
全市政法系统要认真学习领会会议精神，深入领会会议要求，切实增强“四个意识”、坚定“四个自信”、做到“两个维护”，紧紧围绕全市经济社会发展大局，全面加强社会治理，确保人民安居乐业、社会安定有序，为全市经济社会发展作出新的更大贡献。

## 4、中译英测试：

Prompt：把这句话翻译成英文：相濡以沫，不如相忘于江湖。


## Chat-GPT 的回复：

"相濡以沫，不如相忘于江湖"的英文翻译是："To live together with someone, in order to be forgotten in the river of the world, is better than to cherish each other."

## NewBing 的回复：

"相濡以沫，不如相忘于江湖" is a Chinese idiom that means "It's better to forget each other than to help each other in vain." [1](#) [2](#) [3](#)

GPT-4 的回复：

 Translate this sentence into English: It's better to forget each other in the vast world than to comfort each other with tenderness.

## 5、数学推理测试：

Prompt：一把锁的密码，由三个数字组成。请根据下面五个条件推出开锁密码。

- 1.三位数 246，有 1 个号码正确，且位置正确。
- 2.三位数 258，有 1 个号码正确，但位置不正确。
- 3.三位数 692，有 2 个号码正确，但位置都不正确。
- 4.三位数 174，没有一个号码正确。
- 5.三位数 419，有 1 个号码正确，但位置不正确。

Chat-GPT 的回复：



根据条件 1，密码可能是 246、264、462 或 624。根据条件 2，密码可能是 258、285 或 852。根据条件 3，密码可能是 692、962 或 269。根据条件 4，密码可能是 174、141 或 411。根据条件 5，密码可能是 419、149 或 941。

为了确定密码，我们需要验证每个可能性是否满足所有条件。根据条件 1 和 2，我们需要验证是否存在正确的号码和正确的位置。根据条件 3 和 4，我们需要验证是否存在正确的号码，但位置不正确。根据条件 5，我们需要验证是否存在正确的号码，但位置不正确。

对于每个可能性，我们可以列出所有可能的组合，并检查它们是否符合所有条件。例如，对于密码 624，我们可以列出以下组合：624、264、462、246、426、642 和 262。我们检查每个组合是否满足条件 1、2、3、4 和 5，发现只有 624 满足条件 1 和 2，但不满足条件 3 和 4，因此它不是正确的密码。

类似地，对于密码 258，我们可以列出以下组合：258、528、825、582 和 852。我们检查每个组合是否满足条件 1、2、3、4 和 5，发现只有 258 满足条件 1 和 2，但不满足条件 3 和 4，因此它不是正确的密码。

对于密码 246，我们可以列出以下组合：246、624、462、264 和 426。我们检查每个组合是否满足条件 1、2、3、4 和 5，发现只有 246 满足条件 1 和 2，但不满足条件 3 和 4，因此它不是正确的密码。

对于密码 149，我们可以列出以下组合：149、914、491 和 194。我们检查每个组合是否满足条件 1、2、3、4 和 5，发现只有 149 满足条件 1 和 2，但不满足条件 3 和 4，因此它不是正确的密码。

对于密码 941，我们可以列出以下组合：941、491、149 和 194。我们检查每个组合是否满足条件 1、2、3、4 和 5，发现只有 941 满足条件 1 和 2，但不满足条件 3 和 4，因此它不是正确的密码。

因此，正确的密码是 246。

NewBing 的回复：

根据题目中的五个条件，我们可以推出开锁密码是 **042**。下面是推导过程：

- 从第一行得知，246有1个号码正确且位置正确，因此第二个数字是4。
- 从第二行得知，258有1个号码正确但位置不正确，因此第二个数字不是5或8，而是4。
- 从第三行得知，692有2个号码正确但位置都不正确，因此第一个数字是0或2。但由于第四行的信息，0被排除了。因此第一个数字是2。
- 从第四行得知，174没有一个号码正确。因此第三个数字不是1或4。
- 从第五行得知，419有1个号码正确但位置不正确。因此第三个数字是2。

所以开锁密码是042。 <sup>1</sup>

GPT-4 的回复：

🌀 我们可以根据这五个条件逐一分析：

三位数246，有1个号码正确，且位置正确。这意味着密码中可能包含2、4或6。我们暂时无法确定哪个数字是正确的。

三位数258，有1个号码正确，但位置不正确。这意味着密码中可能包含2、5或8，但这个数字的位置与258中的位置不同。我们暂时无法确定哪个数字是正确的。

三位数692，有2个号码正确，但位置都不正确。这意味着密码中可能包含6、9或2，但这两个数字的位置与692中的位置不同。结合条件1，我们可以推断出6是正确的数字，且它的位置是第一位。

三位数174，没有一个号码正确。这意味着密码中不包含1、7和4。结合之前的分析，我们可以确定密码的第二位是2。

三位数419，有1个号码正确，但位置不正确。根据条件4，我们知道1和4都是错误的数字。因此，密码中包含9，但位置不同。所以，密码的第三位是9。

综上所述，一把锁的密码是629。

实验结果

	文学创作	知识问答	新闻撰写	中译英	数学推理	总分
Chat-GPT	2	3	3	1	1	10
NewBing	3	2	2	5	1	13
GPT-4	4	4	4	4	1	17

结论

通过使用相同 prompt 对三个大模型的各方面性能进行测试，可以发现对于简单的文学创作、知识问答三个大模型水平相当，new-bing 由于接入了网络搜索引擎有时可以找到更准确的答案，但推理能力不如 GPT-4，三者 在数学推理题上的表现都不好，说明大模型目前还没有具备完善的数理逻辑，综合来说，GPT-4 的表现最好，是目前应用价值最高的大模型。