目前市面主要AI各大模型对应的版本，收费情况及适合的场景综合对比				
名称	**版本**	**输入收费标准(元/百万汉字)**	**输出收费标准(元/百万汉字)**	**综合应用分析**
Azure OpenAI	GPT 4o	23	92	2024年5月发布，OPENAI目前性能最好的模型，具备较好的推理能力。与GPT4相比，响应速度更快，价格低，并且更擅长处理中文文本，适合各类复杂的任务。
	GPT 4o mini	1	6	2024年7月发布，推理能力介于GPT3.5和GPT4O之间，响应速度快且价格实惠，特别适合高频调用大语言模型的简单业务场景。
	GPT 4	92	252	2023年3月发布，擅长复杂推理，较OPENAI性能最好的模型GPT4O,响应速度慢且价格更贵，建议便用GPT4O。
	GPT 3.5	1	1	2022年11月发布，向公众开放的首款商用大模型，较OPENAI最新的模型GPT4OMINI，推理能力差。
DeepSeek	DeepSeek-V2	2	2	2024年5月发布，擅长通用对话任务，中文表现效果不错
	DeepSeek-Coder- V2	2	2	2024年5月发布，擅长处理编程，数学计算等任务，性价比高
百度文心一言	Ernie-4	198	198	2023年10月发布，百度最强模型，与Ernie-3.5相比，能力全面升级，适用于更加复杂的分析场景，并且这个模型自动对接百度搜索插件，保障问答信息时效。
	Ernie-3.5	20	20	2023年8月发布，对比百度最新的模型Ernie-4，胜在处理带度和价格
	Ernie-Speed-128k	免费	免费	2024年1月发布，百度免费的大语言模型，支持128K上下文窗口，适合需要速度快和处理超长文本的工作，最多一次可输出6100个汉字
	ERNIE-Lite-8K	免费	免费	2023年9月发布，百度免费的大语言模型，响应速度快，但产生的内容质量一般，适合简单任务
	Moonshot-v1-128k	105	105	2024年2月发布，专注于超长文本处理的性价较高的模型之一，上下文窗口支持128K，解析精准度较高，能够很好地理解并按照用户的指示进行操作和生成内容

月之暗面	Moonshot-v1-32k	41	41	2024年2月发布，适合处理中长文本工作，解析精准度高，能够很好地理解并按照用户的指示进行操作和生成内容
	Moonshot-v1-8k	21	21	2024年2月发布，适合处理短文本工作
Claude	Claude V3.5 Sonnet	28	139	2024年6月发布，在模型性能和反应速度之间做了平衡，适合需要同时追求高效和稳定表现的场景。非常适合做超长文字理解和处理的任务
	Claude V3 Haiku	2	12	2024年3月发面，目前是CLAUDE模型家许中反应速度最快的模型，适合非常注重要快速交互的场景，适合做超长言语这也理解和处理的任务
阿里通义千问	qwen-max	48	145	2023年12月发岸上，是通议千问模型家庭中性能最强的模型，适合需要深 入分析和解决复杂问题的场景，但是上下文窗品暂时只支持6K，适合短文本处理任务
	qwen-plus	5	15	2023年10月发布，支持中文，英文等不同语言输入，擅长知识百科，以及进行深入推理的任务
	qwen-turbo	2	7	2023年10月发布，推理能力较弱，价格较便 宜
	qwen 2.5-72b	5	15	2023年9月发布，通知千问目前最强的开源模型，也是目前市面上较强的开源模型，适合需要复杂推理的任务，可处理较长的文本，一次性可输出9000个汉字
	qwen 2.5-32b	4	8	2023年9月发布，通义千问的开源模型，支持本地部署，适合即要较强推理能力又要兼顾反应速度的任务，
	qwen 2.5-14b	2	7	2023年9月发布，通义千问的开源模型，支持本地部署，适合简单任务
清华智谱	GLM4	121	121	2024年1月发布，中文理解能力超过GPT4.适合处理长文本以及复杂推理任务
	GLM-4-Long	1	1	2024年8月发布，支持100万TOKENS的超长上下文窗品，约150-200万字，适用于需要理解大量文本数据的场景，如学术论文，法律文件，历史 文献等的分析总结
	GLM-4-Flash	免费	免费	2024年8月发布，智谱免费的大语言模型，支持128K上下文窗口，适合需要速度快和处理超长文本的工作。

	GLM-3-Turbo	1	1	2023年10月发布，性价比较高，最便宜的模型之一，上下文窗口支持128K，适合推理要求较低的长文本任务；
火山引擎	Doubao-pro-128k	8	8	2024年5月发布，适合处理复杂任务，如参考问答，总结摘要，创作，文本分类和角色扮演等。上下文窗品支持128K，适合处理超长文本的复杂任务场景。
	Doubao-pro-32k	3	3	2024年5月发布，适合处理复杂任务，如参考问答，总结摘要，创作，文本分类和角色扮演等。上下文窗品支持32K，
	Doubao-lite-128k	2	2	2024年5月发布，处理速度较DOUBAO-PRO更愉快，支持 128K的输入输出长度，适合需要快速处理超长文本的场景
	Doubao-lite-32k	1	1	2024年5月发布，处理速度较DOUBAO-PRO更愉快，支持 32K的输入输出长度，适合注重响应度的场景
讯飞星火	Spark4.0 Ultra	121	121	2024年6月发布，目前是讯飞星火模型家族中性能最强的模型，中文对标GPT4—Turbo，适用于文本生成，语言理解，知识问答，逻辑推理，数学能力等复杂任务场景。
	Spark Pro-128K	36	36	2024年7月正式发布，专业级大语言模型，具有百亿级参数，支持128K的输入输出长度，适用于智能问答等对性能，响应速度以及文本处理长度有较高要求的业务场景
	Spark Max	36	36	2024年1月发布，讯飞目前较强的模型之一，中文对标GPT4，适用于数量计算，逻辑推理等对效果有列高要求的业务场景
	Spark Lite	免费	免费	2023年6月正式发布，科大讯飞免费的大语言模型，响应速度快，但是推理能力一般，仅适合简单任务
百川智能	Baichuan4	165	165	2024年5月发布，百川性能最好的模型，此模型融合了海量金融领域专定知识和数据，特别适合金融相关的应用场景
	Baichuan3-Turbo	20	20	2024年1月发布，相较于BAICHUAN2模型，重点优化了内容创作、内容润色，文本分析等企业的高频场景，支持32K的输入输出长度
	Baichuan3-Turbo-128l	40	40	2024年1月发布，相较于BAICHUAN2模型，重点优化了内容创作、内容润色，文本分析等企业的高频场景，支持128K的输入输出长度，适合长文本处理场景
	Baichuan2-Turbo	13	13	2023年12月发布，价格较低的模型之一，但是模型性能一般

	Baichuan2-Turbo-192	26	26	2023年12月发布，支持192K超长上下文窗口，适合做超长文本理解和处理，但是模型性能一般。
商汤日日新	SenseChat-5	48	121	2024年7月发布，商汤目前性能最好的模型，此模型大量使用合成高阶思维链数据，在数理逻辑、英文、指令跟随等方面能力较强
	SenseChat-32K	15	15	2023年4月发布，擅长行业分析和报告撰写，推理能力一般。