模型评测

	选择题（816题）	判断题（2304题）	简题（1019题）
qwen3-max	467/816 = 57.2	1567/2304 = 68.0	68.3
gpt-4o	423/816 = 51.8	820/2304 = 35.6	63.4
gpt-4.1	439/816 = 53.8	854/2304 = 37.1	68.5
o4-mini	460/816 = 56.4	1226/2304 = 53.2	60.3
o4-mini-high	418/816 = 51.2	1192/2304 = 51.7	60.1
Qwen2.5-VL-72B-Instruct	442/816 = 54.2	1215/2304 = 52.7	65.0
Ours	516/816 = **63.2**	1658/2304 = **72.0**	**77.8**

模型名称/学科维度（得分）	考古（2383题）	文物（487题）	历史（619题）	历史地理（234题）	历史文献（231题）	古典文献（69题）	汉语言文字（152题）	古代文学（19题）
qwen3-max	64.1	60.8	71.7	63.4	68.6	73.9	65.6	77.3
gpt-4o	44.8	37.7	46.6	52.5	37.0	40.6	46.0	35.8
gpt-4.1	49.1	42.7	53.7	49.2	31.6	47.8	45.9	40.0
o4-mini	57.4	52.4	62.8	59.1	54.6	57.0	53.6	45.8
o4-mini-high	55.6	48.5	60.6	59.6	50.5	54.4	54.1	47.4
Qwen2.5-VL-72B-Instruct	55.8	54.8	63.0	56.9	50.6	58.4	58.4	60.1
Ours	**66.9**	**72.1**	**76.3**	**76.2**	**68.0**	**74.9**	**72.0**	**61.5**

模型名称/研究场景维度（得	检索（39题）	翻译（62题）	识读（92题）	事实呈现	特征描述	信度考据（78题）	分析推论

分）				（568 题）	（635 题）		（1273 题）
qwen3-max	55.6	72.2	55.0	58.8	58.4	**70.9**	65.3
gpt-4o	55.8	65.9	52.4	54.5	54.1	67.6	61.7
gpt-4.1	59.7	66.9	58.3	57.0	60.5	68.7	63.3
o4-mini	52.2	51.2	56.6	59.1	62.9	62.9	65.2
o4-mini-high	53.9	40.3	56.5	55.6	62.0	61.5	62.2
Qwen2.5-VL-72B-Instruct	54.8	76.6	54.2	56.8	58.8	64.9	60.4
Ours	**62.4**	**80.5**	**65.3**	**70.4**	**63.5**	67.8	**69.9**

模型名称/历史分期维度（得分）	旧石器时代（113 题）	新石器时代（含夏）（1658 题）	商（603 题）	西周（333 题）	东周（492 题）	秦（170 题）	西汉（485 题）	西汉之后（80 题）
qwen3-max	67.2	65.2	64.8	65.6	68.0	71.1	67.9	70.7
gpt-4o	47.3	45.1	48.3	53.8	53.1	58.3	54.4	40.5
gpt-4.1	41.4	45.6	49.0	55.0	57.3	58.3	56.2	38.0
o4-mini	58.6	59.4	58.5	60.3	59.7	60.9	59.4	55.0
o4-mini-high	53.1	59.2	60.3	62.8	61.4	62.9	62.3	55.0
Qwen2.5-VL-72B-Instruct	48.1	53.7	54.1	57.6	59.9	58.7	60.1	51.8
Ours	**86.8**	**68.9**	**70.7**	**70.2**	**73.0**	**72.3**	**72.1**	**75.0**