Measuring economic knowledge

Final evaluation / 12 February 2026

Six language models answered 175 questions on general equilibrium and welfare economics. This public edition presents the overall, topic and difficulty results from the final report.

How the answers were scored

Each question specified a task, an answer format and three economic criteria. Candidates saw the task and format. Gemini 2.5 Flash graded each saved answer, using the rubric to award one point for each criterion met. It also flagged departures from the requested format.

The score is the share of 525 available rubric points. An answer can earn partial credit. The format column counts flagged answers and is reported separately from the points.

Scores from the final grading pass

Candidate model	Points	Point share	Format flags
GPT-5.2	483 / 525	**92.0%**	0
Claude Opus 4.6	473 / 525	**90.1%**	2
Claude Opus 4.5	462 / 525	**88.0%**	4
Gemini 3 Pro Preview	452 / 525	**86.1%**	0
Gemini 2.5 Flash	399 / 525	**76.0%**	0
Gemini 2.0 Flash	307 / 525	**58.5%**	3

How this report relates to the notebook

The teaching notebook revisits an earlier grading pass with Gemini 2.0 Flash. It keeps 154 questions with usable judgments for all three of its selected candidates. The six-model table above uses the final judge and the full set of 175 questions. Both the grading model and the denominator matter when comparing the two analyses.

Where the points came from

Each cell gives the percentage of available rubric points awarded within that group. The item count shows how much evidence each percentage represents.

By topic

Group	Items	GPT-5.2	Opus 4.6	Opus 4.5	Gemini 3 Pro Preview	Gemini 2.5 Flash	Gemini 2.0 Flash
General equilibrium	48	91.7	88.9	88.9	81.9	75.7	61.8
Welfare economics	37	97.3	94.6	93.7	93.7	76.6	63.1
Dynamic models	25	85.3	84.0	80.0	81.3	72.0	50.7
Production	22	95.5	87.9	86.4	81.8	78.8	63.6
Asset pricing	12	94.4	94.4	91.7	91.7	75.0	63.9
Trade	11	97.0	97.0	93.9	97.0	87.9	48.5
Incomplete markets	10	90.0	93.3	86.7	86.7	73.3	43.3
Firms	9	74.1	81.5	74.1	77.8	66.7	51.9
Public economics	1	100.0	100.0	100.0	100.0	100.0	66.7

By difficulty assigned to the question

Group	Items	GPT-5.2	Opus 4.6	Opus 4.5	Gemini 3 Pro Preview	Gemini 2.5 Flash	Gemini 2.0 Flash
Easy	14	88.1	83.3	88.1	81.0	78.6	59.5
Medium	105	89.8	89.5	84.4	83.2	74.9	61.0
Hard	56	97.0	92.9	94.6	92.9	77.4	53.6

Reading a difference in scores

Public economics has one item. A single rubric point there changes the model's topic percentage by 33.3 points. The 48 general-equilibrium items allow 144 points, so a single mark moves that percentage by about 0.7 points.

Every score reflects one judge's application of the rubric to saved answers. Gemini 2.5 Flash also appears among the candidates. A blind human rescore of a sample would help assess how consistently the criteria were applied.