Which answers can you use?

`twin2k_gpt41mini_synthetic.csv against twin2k_wave4_human.csv`

> Precomputed example on public data (Twin-2K-500: human respondents against GPT-4.1-mini synthetic respondents). It is shown for illustration; it is not a verdict on any vendor.

Scope: 87 single-select or ordinal questions scored from 2,058 synthetic and 2,058 human respondents. The two files are compared as answer distributions, not respondent by respondent.

USE SYNTHETIC	SPOT-CHECK WITH HUMANS	FIELD WITH HUMANS
17	23	47
17 from your human data, 0 from the ranker alone	23 from your human data, 0 from the ranker alone	47 from your human data, 0 from the ranker alone

87 questions scored, 91 not scored (multi-select, open text, set aside for review, and similar). 87 have enough human answers to compare directly (human n at least 50); the rest are ordered by the ranker, which ranks and does not give probabilities.Bins marked "ranker only" are provisional.

Not scored: demographic and screening variables (5)

These columns describe respondents or the sample, not survey questions, so they are not audited. Add column names to `config.exclude` to leave out others.

- excluded: demographic/screening variable "gender" is not scored
- excluded: demographic/screening variable "age_band" is not scored
- excluded: demographic/screening variable "region" is not scored
- excluded: demographic/screening variable "education" is not scored
- excluded: demographic/screening variable "party" is not scored

The five riskiest questions

01 Please rate the following technology or products from "not at all risky" to "extremely risky" - bicycles

- Field with humans

Observed gap 0.70 exceeds the 0.15 threshold even after allowing 0.04 for sampling noise (n=2058 human, 2058 synthetic).

02 Would you support or oppose... - Offering seniors healthcare vouchers to purchase private healthcare plans in place of traditional medicare coverage?

- Field with humans

Observed gap 0.70 exceeds the 0.15 threshold even after allowing 0.04 for sampling noise (n=2058 human, 2058 synthetic).

03 Imagine that the U.S. is preparing for the outbreak of an unusual disease, which is expected to kill 600 people. Two alternative programs to combat the disease...

- Field with humans

Observed gap 0.68 exceeds the 0.15 threshold even after allowing 0.05 for sampling noise (n=1003 human, 1003 synthetic).

04 Imagine that a group of research scientists in the School of Medicine are running a laboratory experiment on a vaccine for a rare and fatal virus. The possibil...

- Field with humans

Observed gap 0.66 exceeds the 0.15 threshold even after allowing 0.08 for sampling noise (n=673 human, 673 synthetic).

05 Please rate the following technology or products from "not at all beneficial" to "extremely beneficial" - alcoholic beverages

- Field with humans

Observed gap 0.66 exceeds the 0.15 threshold even after allowing 0.04 for sampling noise (n=2058 human, 2058 synthetic).

If you can only check 10 questions, check these

01 Imagine that highway safety experts have determined that a substantial number of people are at risk of dying in a type of automobile fire. A requirem...

- Spot-check with humans

· borderline: more people will not settle it

02 Would you support or oppose... - A 'public option', which would allow Americans to buy into a government-run healthcare plan if they choose to do so?

- Spot-check with humans

· borderline: more people will not settle it

03 Would you support or oppose... - Immigration reforms that would provide a path to U.S. citizenship for undocumented immigrants currently in the Unite...

- Spot-check with humans

· borderline: more people will not settle it

04 Would you support or oppose... - Federal investments to ensure a carbon-pollution free electricity sector by 2035?

- Spot-check with humans

· borderline: more people will not settle it

05 36. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category: The produ...

- Spot-check with humans

· borderline: more people will not settle it

06 20. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category: The produ...

- Spot-check with humans

· borderline: more people will not settle it

07 33. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category: The produ…

- Spot-check with humans

· borderline: more people will not settle it

08 31. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category: The produ…

- Spot-check with humans

· borderline: more people will not settle it

09 24. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category: The produ…

- Spot-check with humans

· borderline: more people will not settle it

10 32. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category: The produ…

- Spot-check with humans

· borderline: more people will not settle it

Every question

QUESTION	RECOMMENDATION	GAP (TVD)	NOISE	RISK RANK	HUMAN N	SPOT-CHECK N
Please rate the following technology or products from "not at all risky" to "extremely risky" - bicycles	Field with humans	0.70	0.04	11	2,058	borderline
Would you support or oppose... - Offering seniors healthcare vouchers to purchase private healthcare plans in place of traditional medicare...	Field with humans	0.70	0.04	25	2,058	borderline
Imagine that the U.S. is preparing for the outbreak of an unusual disease, which is expected to kill 600 people. Two alternative programs t...	Field with humans	0.68	0.05	2	1,003	borderline
Imagine that a group of research scientists in the School of Medicine are running a laboratory experiment on a vaccine for a rare and fatal...	Field with humans	0.66	0.08	1	673	borderline
Please rate the following technology or products from "not at all beneficial" to "extremely beneficial" - alcoholic beverages	Field with humans	0.66	0.04	4	2,058	borderline
Imagine that the U.S. is preparing for the outbreak of an unusual disease, which is expected to kill 600 people. Two alternative programs t...	Field with humans	0.63	0.06	9	1,055	borderline
Imagine that when you went to the movies last week, you were inadvertently exposed to a rare and fatal virus. The possibility of actually c...	Field with humans	0.61	0.06	3	673	borderline
Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of...	Field with humans	0.60	0.05	14	1,029	borderline
Please rate the following technology or products from "not at all risky" to "extremely risky" - alcoholic beverages	Field with humans	0.58	0.04	23	2,058	borderline
You have recently graduated from university, obtained a good job, and are buying a new car. A newly designed seatbelt has just become avail...	Field with humans	0.57	0.06	6	735	borderline
Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of...	Field with humans	0.57	0.06	47	1,029	borderline
Please rate the following technology or products from "not at all beneficial" to "extremely beneficial" - pesticides	Field with humans	0.56	0.04	20	2,058	borderline

QUESTION	RECOMMENDATION	GAP (TVD)	NOISE	RISK RANK	HUMAN N	SPOT-CHECK N
Imagine that there will be a deadly flu going around your area next winter. Your doctor says that you have a 10% chance (10 out of 100) of...	Field with humans	0.53	0.03	10	2,058	borderline
Please rate the following technology or products from "not at all risky" to "extremely risky" - chemical plants	Field with humans	0.52	0.03	17	2,058	borderline
Please rate the following technology or products from "not at all beneficial" to "extremely beneficial" - chemical plants	Field with humans	0.51	0.04	13	2,058	borderline
A 55-year-old man had a heart condition. He had to stop working because of chest pain. He enjoyed his work and did not want to stop. His pa...	Field with humans	0.50	0.04	15	1,006	borderline
You have recently graduated from university, obtained a good job, and are buying a new car. A newly designed seatbelt has just become avail...	Field with humans	0.50	0.07	7	651	borderline
You have recently graduated from university, obtained a good job, and are buying a new car. A newly designed seatbelt has just become avail...	Field with humans	0.50	0.07	12	672	borderline
Please rate your level of disagreement or agreement with the following statement: "I would find a game that had a 7/36 chance of winning $9...	Field with humans	0.49	0.07	34	672	borderline
Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of...	Field with humans	0.48	0.05	21	1,029	borderline
Please rate the following technology or products from "not at all risky" to "extremely risky" - pesticides	Field with humans	0.48	0.04	24	2,058	borderline
Imagine that when you went to the movies last week, you were inadvertently exposed to a rare and fatal virus. The possibility of actually c...	Field with humans	0.47	0.08	29	712	borderline
Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of...	Field with humans	0.47	0.05	30	1,029	borderline
A 55-year-old man had a heart condition. He had to stop working because of chest pain. He enjoyed his work and did not want to stop. His pa...	Field with humans	0.46	0.04	16	1,052	borderline

QUESTION	RECOMMENDATION	GAP (TVD)	NOISE	RISK RANK	HUMAN N	SPOT-CHECK N
According to a comprehensive study by the U.S. Department of Transportation, a particular German car is 8 times more likely than a typical...	Field with humans	0.46	0.06	5	1,043	borderline
Please rate your level of disagreement or agreement with the following statement: "I would find a game that had a 7/36 chance of winning $9...	Field with humans	0.45	0.07	41	735	borderline
According to a comprehensive study by the U.S. Department of Transportation, Ford Explorers are 8 times more likely than a typical family c...	Field with humans	0.43	0.06	8	1,015	borderline
Would you support or oppose... - A 2% tax on the assets of individuals with a net worth of more than $50 million?	Field with humans	0.43	0.04	44	2,058	borderline
Please rate your level of disagreement or agreement with the following statement: "I would find a game that had a 7/36 chance of winning $9...	Field with humans	0.41	0.07	43	651	borderline
Would you support or oppose... - Increasing deportations for those in the US illegally?	Field with humans	0.40	0.04	38	2,058	borderline
Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of...	Field with humans	0.39	0.06	28	1,029	borderline
Would you support or oppose... - Placing a tax on carbon emissions?	Field with humans	0.34	0.04	32	2,058	borderline
Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of...	Field with humans	0.29	0.05	33	1,029	borderline
Please rate the following technology or products from "not at all beneficial" to "extremely beneficial" - bicycles	Field with humans	0.29	0.03	45	2,058	borderline
Imagine that highway safety experts have determined that a substantial number of people are at risk of dying in a type of automobile fire....	Field with humans	0.26	0.06	26	735	borderline
Would you support or oppose... - A 'Medicare for All' system in which all Americans would get healthcare from a government-run plan?	Field with humans	0.23	0.04	37	2,058	borderline
Imagine that you go to purchase a calculator for $30. The calculator salesperson informs you that the calculator you wish to buy is on sale...	Field with humans	0.23	0.03	19	1,031	borderline

QUESTION	RECOMMENDATION	GAP (TVD)	NOISE	RISK RANK	HUMAN N	SPOT-CHECK N
Imagine that highway safety experts have determined that a substantial number of people are at risk of dying in a type of automobile fire....	Field with humans	0.22	0.06	18	672	borderline
Assume that you are presented with two trays of black and white marbles, a large tray that contains 100 marbles and a small tray that conta...	Field with humans	0.22	0.03	39	2,058	borderline
38. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Field with humans	0.22	0.03	48	2,058	borderline
40. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Field with humans	0.21	0.03	50	2,058	borderline
30. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Field with humans	0.20	0.03	53	2,058	borderline
28. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Field with humans	0.19	0.03	49	2,058	borderline
34. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Field with humans	0.19	0.03	52	2,058	borderline
29. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Field with humans	0.19	0.03	54	2,058	borderline
35. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Field with humans	0.18	0.03	55	2,058	borderline
39. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Field with humans	0.18	0.03	51	2,058	borderline
Would you support or oppose... - A ‘public option’, which would allow Americans to buy into a government-run healthcare plan if they choose...	Spot-check with humans	0.18	0.04	35	2,058	borderline
31. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Spot-check with humans	0.17	0.03	59	2,058	borderline

QUESTION	RECOMMENDATION	GAP (TVD)	NOISE	RISK RANK	HUMAN N	SPOT-CHECK N
36. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Spot-check with humans	0.17	0.03	56	2,058	borderline
20. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Spot-check with humans	0.17	0.03	57	2,058	borderline
33. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Spot-check with humans	0.17	0.03	58	2,058	borderline
24. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Spot-check with humans	0.17	0.03	60	2,058	borderline
18. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Spot-check with humans	0.16	0.03	64	2,058	borderline
Imagine that highway safety experts have determined that a substantial number of people are at risk of dying in a type of automobile fire....	Spot-check with humans	0.16	0.06	22	651	borderline
32. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Spot-check with humans	0.16	0.03	61	2,058	borderline
22. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Spot-check with humans	0.16	0.03	63	2,058	borderline
21. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Spot-check with humans	0.15	0.03	66	2,058	borderline
Would you support or oppose... - Immigration reforms that would provide a path to U.S. citizenship for undocumented immigrants currently in...	Spot-check with humans	0.14	0.04	36	2,058	borderline
26. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Spot-check with humans	0.14	0.03	67	2,058	borderline
11. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Spot-check with humans	0.14	0.03	68	2,058	borderline

QUESTION	RECOMMENDATION	GAP (TVD)	NOISE	RISK RANK	HUMAN N	SPOT-CHECK N
15. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Spot-check with humans	0.14	0.03	72	2,058	borderline
14. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Spot-check with humans	0.13	0.03	62	2,058	borderline
10. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Spot-check with humans	0.13	0.03	75	2,058	borderline
12. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Spot-check with humans	0.13	0.03	71	2,058	borderline
8. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Spot-check with humans	0.13	0.03	70	2,058	borderline
9. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Spot-check with humans	0.13	0.03	65	2,058	borderline
25. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Spot-check with humans	0.13	0.03	73	2,058	borderline
1. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Spot-check with humans	0.13	0.03	74	2,058	borderline
Would you support or oppose... - Federal investments to ensure a carbon-pollution free electricity sector by 2035?	Spot-check with humans	0.08	0.04	46	2,058	borderline
5. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Use synthetic	0.12	0.03	80	2,058	–
6. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Use synthetic	0.12	0.03	78	2,058	–
16. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Use synthetic	0.12	0.03	79	2,058	–
37. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Use synthetic	0.11	0.03	77	2,058	–

QUESTION	RECOMMENDATION	GAP (TVD)	NOISE	RISK RANK	HUMAN N	SPOT-CHECK N
23. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Use synthetic	0.11	0.03	69	2,058	–
19. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Use synthetic	0.11	0.03	82	2,058	–
7. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Use synthetic	0.10	0.03	84	2,058	–
27. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Use synthetic	0.10	0.03	76	2,058	–
13. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Use synthetic	0.10	0.03	81	2,058	–
17. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Use synthetic	0.10	0.03	83	2,058	–
Would you support or oppose... - Ensuring 40% of all new clean energy infrastructure development spending goes to low-income communities?	Use synthetic	0.09	0.04	40	2,058	–
Would you support or oppose... - A law that requires companies to provide paid family leave for parents?	Use synthetic	0.08	0.04	27	2,058	–
Imagine that you go to purchase a jacket for $250. The jacket salesperson informs you that the jacket you wish to buy is on sale for $240 a...	Use synthetic	0.07	0.04	42	1,027	–
4. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Use synthetic	0.07	0.03	85	2,058	–
3. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Use synthetic	0.06	0.03	86	2,058	–
2. Please consider the following product category. Suppose you are in a grocery store, and you see the following product in that category:...	Use synthetic	0.05	0.03	87	2,058	–
Do you think there are more or fewer than 65 African countries in the United Nations?	Use synthetic	0.02	0.03	31	1,056	–

Where it breaks by subgroup

Each cell is one question within one subgroup level. Cells with too few human answers are labelled too small to judge. Treat single cells as leads, not findings.

SUBGROUP	LEVEL	DIVERGES	UNCLEAR	MATCHES	TOO SMALL	MEAN GAP
age_band	18-29	41	41	5	0	0.29
age_band	30-49	39	40	8	0	0.29
age_band	50-64	37	40	10	0	0.29
age_band	65+	35	49	3	0	0.29
gender	Female	40	27	20	0	0.28
gender	Male	50	29	8	0	0.30
party	Democrat	43	30	14	0	0.31
party	Independent	41	34	12	0	0.30
party	Republican	50	36	1	0	0.34
party	Something else	11	48	1	27	0.21

Largest subgroup gaps

QUESTION	SUBGROUP	HUMAN N	GAP (TVD)	NOISE
Imagine that the U.S. is preparing for the outbreak of an unusual disease, which is expected to kill 600 peop…	party: Republican	285	0.87	0.10
Please rate the following technology or products from "not at all risky" to "extremely risky" - bicycles	party: Something else	62	0.77	0.24
Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was d…	party: Democrat	431	0.77	0.09
Would you support or oppose… - Offering seniors healthcare vouchers to purchase private healthcare plans in…	age_band: 18-29	388	0.77	0.09
Please rate the following technology or products from "not at all risky" to "extremely risky" - bicycles	party: Republican	540	0.76	0.08
Would you support or oppose… - Offering seniors healthcare vouchers to purchase private healthcare plans in…	gender: Female	1,044	0.75	0.05

QUESTION	SUBGROUP	HUMAN N	GAP (TVD)	NOISE
Please rate the following technology or products from "not at all beneficial" to "extremely beneficial" - alc…	party: Independent	609	0.74	0.07
Would you support or oppose… - Offering seniors healthcare vouchers to purchase private healthcare plans in…	party: Democrat	847	0.73	0.07
Would you support or oppose… - Offering seniors healthcare vouchers to purchase private healthcare plans in…	party: Independent	609	0.73	0.07
Would you support or oppose… - Offering seniors healthcare vouchers to purchase private healthcare plans in…	age_band: 30-49	735	0.73	0.07
Please rate the following technology or products from "not at all risky" to "extremely risky" - bicycles	age_band: 50-64	658	0.72	0.07
Imagine that a group of research scientists in the School of Medicine are running a laboratory experiment on…	age_band: 65+	93	0.72	0.24
Please rate the following technology or products from "not at all risky" to "extremely risky" - bicycles	age_band: 65+	277	0.72	0.11
Imagine that a group of research scientists in the School of Medicine are running a laboratory experiment on…	gender: Female	362	0.72	0.11
Imagine that the U.S. is preparing for the outbreak of an unusual disease, which is expected to kill 600 peop…	age_band: 30-49	354	0.71	0.09
Imagine that the U.S. is preparing for the outbreak of an unusual disease, which is expected to kill 600 peop…	party: Independent	302	0.71	0.09
Imagine that a group of research scientists in the School of Medicine are running a laboratory experiment on…	party: Republican	188	0.71	0.16
Please rate the following technology or products from "not at all beneficial" to "extremely beneficial" - alc…	gender: Male	1,014	0.71	0.06
Please rate the following technology or products from "not at all risky" to "extremely risky" - bicycles	gender: Male	1,014	0.71	0.05
Imagine that a group of research scientists in the School of Medicine are running a laboratory experiment on…	party: Democrat	274	0.70	0.13
Please rate the following technology or products from "not at all risky" to "extremely risky" - bicycles	gender: Female	1,044	0.70	0.05
Imagine that the U.S. is preparing for the outbreak of an unusual disease, which is expected to kill 600 peop…	age_band: 50-64	305	0.70	0.10
Please rate the following technology or products from "not at all risky" to "extremely risky" - bicycles	party: Independent	609	0.70	0.07

QUESTION	SUBGROUP	HUMAN N	GAP (TVD)	NOISE
Please rate the following technology or products from "not at all risky" to "extremely risky" - bicycles	age_band: 30-49	735	0.69	0.06
Imagine that the U.S. is preparing for the outbreak of an unusual disease, which is expected to kill 600 peop…	gender: Male	490	0.69	0.07
Would you support or oppose… - A 2% tax on the assets of individuals with a net worth of more than $50 mill…	party: Republican	540	0.68	0.07
Please rate the following technology or products from "not at all beneficial" to "extremely beneficial" - alc…	age_band: 30-49	735	0.68	0.07
Imagine that the U.S. is preparing for the outbreak of an unusual disease, which is expected to kill 600 peop…	age_band: 18-29	198	0.68	0.12
Imagine that a group of research scientists in the School of Medicine are running a laboratory experiment on…	age_band: 50-64	210	0.68	0.15
Please rate the following technology or products from "not at all risky" to "extremely risky" - bicycles	age_band: 18-29	388	0.68	0.09

How the ranker fares on your data

On 87 questions where your human data allows a direct comparison, 67% exceed the failure threshold. The ranker, which sees only the synthetic answers, orders them with an AUC of 0.91 (0.5 is chance, 1.0 is perfect) and the top fifth of its ranking contains 1.50 times the failure rate of a random fifth. With your human answers added as an audit sample, AUC is 1.00.

> If your files contain the Twin-2K-500 study, the figure above is in-sample for the ranker and optimistic. The out-of-sample figures are below.

PUBLISHED EVIDENCE FOR THE RANKER	ROC AUC
Twin-2K, synthetic answers only, question families and respondents held out	0.802
Twin-2K, plus a 10% human audit	0.883
GSS, frozen Twin-2K model, no refit (10 question families)	0.752 [0.554, 0.893]

The ranking transferred to a second study; its probabilities did not (calibration error 0.178 on GSS). This report therefore uses ranks and never shows a probability.

Method and limits

What was done

- Each question's answer options were aligned between the two files. Only matches the intake module was confident about were scored; the others are listed above as set aside.
- Gap: total variation distance (TVD) between the human and synthetic answer distributions. The failure threshold is 0.15. Mean gap is shown for ordered options only (questions.csv).
- Noise allowance: the 95th percentile of the TVD you would see between two samples of your sizes drawn from one and the same distribution (300 simulations, effective n when weighted, seed 7). A question is called a failure only when its gap exceeds the threshold after subtracting this allowance, and called safe only when the gap plus the allowance stays below it. Everything else is a spot-check.
- Cells with fewer than 50 effective human answers are labelled too small to judge.
- Ranker: a logistic regression frozen on 212 question cells from 75 question families of the public Twin-2K-500 study (revision f883165a). It sees only the shape of the synthetic answer distribution (and, for the audit variant, a human sample).
- Questions without enough human answers are binned by their ranker rank: the riskiest fifth to "Field with humans", the next two fifths to "Spot-check with humans", the rest to "Use synthetic", marked provisional. These are forced proportions, not an estimate of how many questions diverge.
- Demographic and screening columns (age, gender, region, education, income, race, party, weights, IDs) and any column in `config.exclude` are not scored.
- Spot-check sizes are the human sample at which the expected sampling noise reaches 0.05.

What this does not tell you

- The ranker rests on one dataset, one model release and one questionnaire, plus one transfer test on ten question families. It is exploratory evidence, and a ranking, not a probability.
- The 0.15 gap is a review trigger, not a significance test and not a business loss.
- Answers are compared as most-common-answer and distribution only. Real decisions need a loss function this report does not model.
- Survey weights are used for the observed gaps and effective n. Design effects beyond the Kish approximation are not modelled.
- Only single-select and ordinal items are supported. Multi-select, open-ended and slider items are not scored.
- Subgroup cells are many comparisons on a finite sample: expect some to diverge by chance.
- This is not a certification of any synthetic vendor.

Ranker training data: LLM-Digital-Twin/Twin-2K-500, licence CC BY 4.0. Report generated by the Kernsignal API.