Conclusion Fidelity of AI Agents in Synthetic Omics Environments

Arinbjörn Kolbeinsson
K01
Reykjavík, Iceland
arinbjorn@K01.is

Benedikt Kolbeinsson
K01
Reykjavík, Iceland
benedikt@K01.is

Abstract

Synthetic data is proposed as a basis for discovery in biomedicine, and AI agents are beginning to carry out the analysis. Synthetic data is accepted on the strength of fidelity metrics, which measure its similarity to real data. However, similarity does not establish that an agent analysing synthetic data reaches the conclusions it would reach on the real data. We build synthetic molecular breast cancer cohorts into which we introduce flaws of known location and strength, ask agents built on three open-weight models to analyse each cohort without a fixed plan, and check every claim they make against the real cohort. Moderately imperfect generators yield claims that almost all hold on the real cohort but miss real findings. A weakened true relationship disappears from the agents' conclusions rather than being reversed, while a fabricated signal is reported as a marker in most runs once it approaches the strength of genuine markers. When a fabrication contradicts textbook biology, the agent notices the conflict but rarely rejects it. Standard fidelity metrics register neither flaw. Re-evaluating each claim on an ensemble of generators separates claims that hold on real data from artefacts, unless all generators share the flaw.

1 Introduction

Synthetic data is offered as a way to work with sensitive biomedical data without giving access to the patient records themselves [Yoon et al., 2020, Chen et al., 2021, Jordon et al., 2022, Gonzales et al., 2023, Giuffrè and Shung, 2023]. Patient-level molecular data is among the most restricted, and synthetic versions of such cohorts could open them to far more researchers. Increasingly, this research is carried out by AI agents that plan analyses, write code and report findings [Boiko et al., 2023, M. Bran et al., 2024, Lu et al., 2024, Swanson et al., 2025]. Before such an agent is used to analyse a synthetic cohort, we need to know whether its conclusions would hold on the real data. As AI accelerates computational research, the bottleneck of discovery is shifting towards validation, and such a check is part of the verification infrastructure that Imas and Manyika [2026] call for.

Synthetic data is usually accepted on the strength of fidelity metrics: distances between real and synthetic distributions such as the maximum mean discrepancy (MMD) [Gretton et al., 2012], or sample-level precision and recall [Alaa et al., 2022]. Such metrics measure similarity, but they do not measure whether the particular relationships an analysis depends on are preserved. A synthetic cohort can therefore be similar to the real one overall while misrepresenting the specific relationships an analysis would report as findings. For example, synthetic clinical time series can improve in conditional MMD while still failing to reproduce the relationship between fever and tachycardia [Kolbeinsson and Kolbeinsson, 2025].

Here, we investigate whether an agent analysing synthetic data reaches the conclusions it would reach on the real data that the synthetic data came from, which we call conclusion fidelity, and whether conclusions that do not hold can be identified. Specifically, the contributions of this paper are:

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: Agentic AI for Biological Discovery.

- A testbed in which flaws of controlled location and strength are introduced into synthetic data and every claim an agent makes is scored automatically against the real cohort (Section 2).
- Evidence that moderately imperfect generators yield discoveries with high precision but lower recall, while one that fails to learn the data yields false ones (Section 3.1).
- How conclusions change with flaw strength: weakening a true relationship lowers recall, while a fabricated one is reported above a threshold set by genuine markers (Section 3.2).
- A score for each claim, based on whether it also holds in data from other generators, and its failure when all generators share a flaw (Section 3.3).

Related work. Previous work has asked whether conclusions hold on synthetic data by rerunning fixed, pre-specified analyses [Rosenblatt et al., 2023, El Emam et al., 2024], and naive inference on synthetic data is known to produce false discoveries [Decruyenaere et al., 2024, Montoya Perez et al., 2024]. In our setting the agent chooses the analysis, and the flaws are introduced at known locations with controlled strength. The closest work, Rewolinski et al. [2026], varies signal strength and uses null data to test the yes/no conclusions of one closed model. Although we share the idea of introducing signals of known strength, we score every claim that open-weight models report on an omics cohort, including a fabrication that contradicts textbook biology. See also Appendix A.

2 Setup

Cohort. We use the TCGA breast cancer cohort [Cancer Genome Atlas Network, 2012] as distributed by UCSC Xena [Goldman et al., 2020]: RNA-seq expression ($\log_2$ normalised counts), PAM50 subtype calls [Parker et al., 2009], receptor status, stage, histology, age, and gene-level somatic mutation calls [Ellrott et al., 2018] for five frequently mutated genes. We restrict the cohort to the 844 primary tumours that have a PAM50 subtype call and to 219 genes: 69 with an established role in breast cancer and 150 little-studied genes (Appendix B).

World models. We call each generator a world model, and the synthetic cohort it produces the agent's world, or environment. Our baseline generator is a Gaussian copula fitted within each PAM50 subtype, which keeps each subtype's marginals and rank correlations [Patki et al., 2016]. We also train two off-the-shelf deep tabular generators, CTGAN and TVAE [Xu et al., 2019], with default settings. We introduce flaws as deterministic edits to the sampled data, so an edited cohort differs from the baseline only in the edited gene. Each edit has a strength parameter that we vary in steps. The *reversal edit* replaces a gene's standardised value z_t by $(z_t - \lambda\rho z_s)/\sqrt{1 - 2\lambda\rho^2 + \lambda^2\rho^2}$, where ρ is its correlation with a source gene z_s: $\lambda = 0$ leaves the data unchanged, $\lambda = 1$ removes the correlation and $\lambda = 2$ reverses it, keeping the gene's mean and variance. We apply it to GRB7, which is co-amplified with ERBB2 and a textbook marker of HER2-enriched tumours. The *fabrication edit* adds d SD to the expression of an unrelated, little-studied gene (C8orf84) in HER2-enriched tumours; in the real cohort, its difference between HER2-enriched and other tumours is -0.09 SD.

Agents. We evaluate three open-weight models from different families, each served locally with vLLM [Kwon et al., 2023] and given access to a Python tool. Qwen3.6-35B-A3B [Qwen Team, 2026, 2025], with its reasoning mode ('thinking') enabled, is used in all experiments. Gemma-4-26B [Gemma Team, 2026] and Nemotron-3-Super-120B [NVIDIA, 2026] are evaluated on a subset of the conditions, and all three are used to evaluate the trust signal (Section 3.3). Unless stated otherwise, results are reported for Qwen3.6. The agents receive a research brief, the cohort as a CSV and a data dictionary, and are not told whether the data are real or synthetic. They choose their own analyses and finish with a written report and a list of structured claims: differences between a group and the rest, or associations between two variables. Two briefs concern HER2-enriched tumours (one asks for candidate biomarkers or drug targets; the other, phrased neutrally, asks how expression differs) and one concerns basal-like tumours. We run each combination of brief and world, a cell, with 8–32 seeds; a seed produces the same first analysis step in every world, so runs are paired.

Scoring. Claims are scored automatically, without an LLM judge: a claim holds if its standardised effect (Cohen's d or Spearman ρ) has the stated direction and clears a fixed threshold ($|d| \geq 0.2$, $|\rho| \geq 0.1$). We test each claim on the real cohort, on the synthetic sample the agent analysed and on a large sample from the same generator. A claim that holds on the real cohort is *correct*. One that

fails there but holds in the agent's sample is a *world artefact* if it also holds in the large sample and a *sampling artefact* if not. One that fails even in the agent's sample is an *agent error*. Here, correct and true mean agreement with the real cohort, an empirical reference rather than causal truth. We summarise conclusions by claim-level precision, the share of claims that are correct, and recall, the share of real findings that are reported (Figure 2).

3 Results

3.1 Discoveries in imperfect world models

Table 1: Conclusions by world model, eight runs each (HER2 targets; basal in brackets).

World	MMD	Corr. err.	Precision	Recall
Real data	–	–	0.99 (1.00)	0.76 (0.64)
Copula	0.001	0.011	0.96 (1.00)	0.59 (0.54)
TVAE	0.016	0.070	0.99 (1.00)	0.49 (0.43)
CTGAN	0.156	0.202	0.54 (0.44)	0.10 (0.01)

Table 1 compares the world models. In the table, correlation error (Corr. err.) is the mean absolute difference of gene–gene correlations from the real cohort, and recall is over the core findings, those claimed in at least half of the real-data runs (Appendix B). Although the copula and TVAE differ by an order of magnitude in MMD to the real cohort, nearly all claims the agent makes on either synthetic cohort hold on the real data, so precision is high. However, recall is lower: the share of core findings reported falls from 0.76 on real data to 0.59 on the copula and 0.49 on TVAE. With eight runs per cell these differences are uncertain: resampling whole runs gives 95% intervals of 0.60–0.89, 0.45–0.70 and 0.36–0.61 (Table 3). Many false claims appear only when a generator fails to learn the data's structure, as CTGAN did with default settings: it lost all gene–gene correlations and subtype differences. Tuning may avoid this. Results on the basal-like brief agree.

3.2 Weakened and fabricated signals

Figure 1a shows both edits on one scale: the edited gene's effect in the synthetic data minus its effect in the real cohort. As the correlation between GRB7 and ERBB2 is weakened and then reversed, recall of GRB7 falls: the agent reports it as a HER2 marker in 12 of 16 runs at $\lambda = 0$, in 2–5 of 8 at $\lambda = 0.5$–1, and in 1 of 24 for $\lambda \geq 1.25$. On the real cohort, and on its bootstrap resamples, it does so in 7 of 8 runs. GRB7 starts to disappear from the reports while it is still strongly enriched in the synthetic data ($d = 1.2$ at $\lambda = 0.75$), because the agent reports a short ranked list of findings, from which a weakened signal drops out early. At full reversal the agent mostly omits GRB7 and states the false direction in only 1 of 8 runs; the neutral brief gives the same result.

A fabricated signal behaves differently. When an unrelated, little-studied gene is given an artificial difference in HER2-enriched tumours, the agent almost never reports it below about 1.3 SD (2 of 48 runs), but from 1.5 SD onwards it reports the gene as a HER2 marker in 20 of 32 runs, against 0 of 16 on the unedited copula. No run describes it as unexpected or anomalous (0 of 72 runs with a fabricated effect). The threshold depends on the strength of the genuine markers. On the basal-like brief, whose genuine markers reach $|d| \approx 2.4$, the same fabrication is reported in 1 of 8 runs at $d = 1.74$ and in 5 of 8 at $d = 2.08$, and reversing the genuine marker FOXC1 removes it from all 8 reports. The false marker also displaces real findings, reducing recall of the core findings from 0.59 to 0.40. Standard fidelity metrics do not register the edits, since for every edited world the correlation error is 0.011–0.013 and the MMD at most 0.002, as for the copula.

Biology. As little-studied genes give the agent nothing to check a fabrication against, we also fabricate a signal that contradicts textbook biology: ESR1, which the agent reports as lower in HER2-enriched tumours in 23 of 24 runs on clean data, is shifted to the same synthetic effect as the little-studied gene ($d = +1.69$). Across both HER2 briefs the agent reports ESR1 as higher in 29 of 32 runs. It notices the conflict in 23, calling the result "unexpected" or "contrary to the clinical paradigm", but in 21 of those it reports the finding anyway, typically attributing it to characteristics of the cohort. It rejects the finding in only two runs. This suggests that the agent's prior knowledge leads it to notice the conflict but rarely to reject the finding.

Other agents. Gemma-4-26B behaves like Qwen3.6: it drops GRB7 under full reversal in all 8 runs and reports the fabricated gene in 5 of 8 without comment. Nemotron-3-Super-120B relies more on its prior knowledge. Fewer of its claims are supported by the data it analysed (faithfulness 0.87–0.95 against 0.99–1.00), under full reversal it asserts the textbook direction for GRB7 in 2 of 15 runs, and it reports the fabrications less often (6 of 16; ESR1 2 of 8), mostly by omitting the gene.

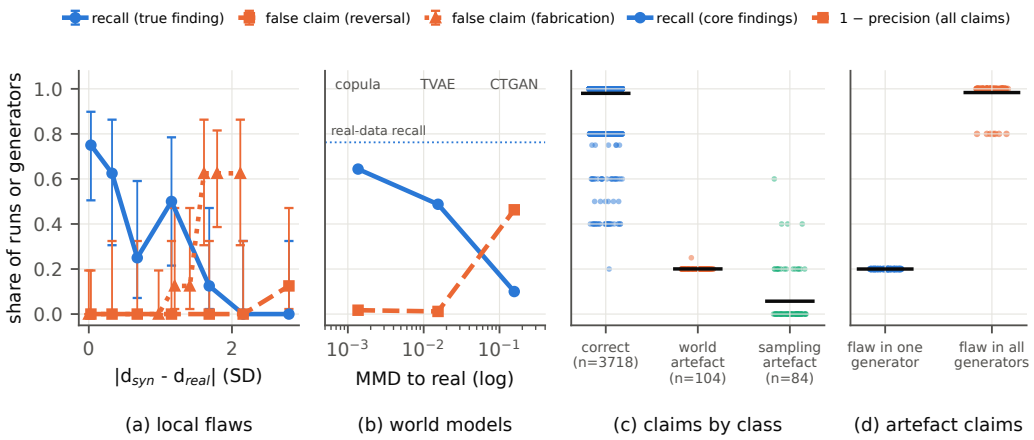


Figure 1: (a, b) Recall (blue) and false claims (orange: runs with a false claim about the manipulated signal, or 1−precision over all claims) against distance from real data: (a) synthetic minus real effect of the manipulated signal (Qwen3.6), (b) MMD. Wilson 95% intervals. (c) Generator agreement per claim, by class, all models and briefs. (d) Artefact claims, flaw in one or all generators.

3.3 Identifying which conclusions to trust

A data holder can fit several generators. We re-evaluate each claim in the agent's generator, TVAE and three bootstrap refits of the copula, and define its agreement as the share of these generators in which it holds. Across all models and briefs, agreement separates the 3,718 correct claims from the 188 world and sampling artefacts with AUROC 0.9998, and 0.984 without the copula refits (the agent's generator, TVAE and CTGAN). Claims from one run are not independent, and a claim repeated on the same world receives the same agreement. Counting each of the 1,089 distinct claims once, or resampling whole runs, leaves the AUROC at 0.999 or above (0.98 without the refits; Table 3). However, agreement cannot detect a flaw that every generator shares (Figure 1d). For the 119 claims about an edited signal, it separates the 85 false from the 34 true perfectly when only one generator has the flaw (0.20 against 0.89, AUROC 1.00), but not when all do (0.98 against 0.88, AUROC 0.46).

When the location of a flaw is known, few claims need checking on real data. For Qwen3.6, checking on real data only the 5–13% of claims that involve an edited gene catches 78–97% of the false claims in worlds with a fabricated signal, and raises precision from 0.93–0.94 to 0.98–1.00. A generator that fails globally, as CTGAN did, cannot be handled in this way, although standard metrics detect it.

4 Discussion

In our setting, standard metrics were sufficient to reject a generator that failed to learn the data's structure, but missed the local flaws that decide individual conclusions. A fabricated signal is more dangerous: above a moderate strength the agent reports it, without comment when the gene is little studied and with an explanation when it contradicts the agent's knowledge. This suggests that findings an agent itself describes as unexpected should be checked on real data before they are reported. Disagreement between independently built generators reveals flaws specific to one of them, while flaws they share can only be found by checking some claims on real data.

Limitations. We use one cohort, three agent models and two kinds of introduced flaw, so our thresholds may not transfer, and 8–32 runs per cell leave wide intervals (Table 3). The trust signal assumes that the generators' flaws are independent, which bootstrap refits of a good copula favour. The testbed measures conclusion fidelity under known flaws and does not certify a generator. Open problems are detecting shared flaws without real data, locating flaws as precisely as when their location is known, and agents that reject implausible findings rather than explain them away.

Acknowledgments and Disclosure of Funding

We thank the University of Iceland and the University of Iceland Science Park for lending the hardware used for all agent runs.

References

Ahmed M. Alaa, Boris van Breugel, Evgeny Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 290–306, 2022.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023. doi: 10.1038/s41586-023-06792-0.

Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012. doi: 10.1038/nature11412.

Richard J. Chen, Ming Y. Lu, Tiffany Y. Chen, Drew F. K. Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497, 2021. doi: 10.1038/s41551-021-00751-8.

Alexander Decruyenaere, Heidelinde Dehaene, Paloma Rabaey, Christiaan Polet, Johan Decruyenaere, Stijn Vansteelandt, and Thomas Demeester. The real deal behind the artificial appeal: Inferential utility of tabular synthetic data. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 244 of *Proceedings of Machine Learning Research*, pages 966–996, 2024.

Khaled El Emam, Lucy Mosquera, Xi Fang, and Alaa El-Hussuna. An evaluation of the replicability of analyses using synthetic health data. *Scientific Reports*, 14:6978, 2024. doi: 10.1038/s41598-024-57207-7.

Kyle Ellrott, Matthew H. Bailey, Gordon Saksena, Kyle R. Covington, Cyriac Kandoth, Chip Stewart, Julian Hess, Singer Ma, Kami E. Chiotti, Michael McLellan, et al. Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Systems*, 6(3): 271–281, 2018. doi: 10.1016/j.cels.2018.03.002.

Gemma Team. Gemma 4 technical report, 2026.

Mauro Giuffrè and Dennis L. Shung. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *npj Digital Medicine*, 6:186, 2023. doi: 10.1038/s41746-023-00927-3.

Mary J. Goldman, Brian Craft, Mim Hastie, Kristupas Repečka, Fran McDade, Akhil Kamath, Ayan Banerjee, Yunhai Luo, Dave Rogers, Angela N. Brooks, Jingchun Zhu, and David Haussler. Visualizing and interpreting cancer genomics data via the Xena platform. *Nature Biotechnology*, 38:675–678, 2020. doi: 10.1038/s41587-020-0546-8.

Aldren Gonzales, Guruprabha Guruswamy, and Scott R. Smith. Synthetic data in health care: A narrative review. *PLOS Digital Health*, 2(1):e0000082, 2023. doi: 10.1371/journal.pdig.0000082.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Ken Gu, Ruoxi Shang, Ruien Jiang, Keying Kuang, Richard-John Lin, Donghe Lyu, Yue Mao, Youran Pan, Teng Wu, Jiaqian Yu, Yikun Zhang, Tianmai M. Zhang, Lanyi Zhu, Mike A. Merrill, Jeffrey Heer, and Tim Althoff. BLADE: Benchmarking language model agents for data-driven science. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13936–13971, 2024.

Alex Imas and James Manyika. Bending the curve of discovery: AI in science today and tomorrow. DeepMind Institute essay, October 2026. URL `https://institute.deepmind.com/essays/bending-the-curve-of-discovery-ai-in-science-today-and-tomorrow/`.

James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N. Cohen, and Adrian Weller. Synthetic data – what, why and how? *arXiv preprint arXiv:2205.03257*, 2022.

Elliot Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2506.07962.

John C. Knight and Nancy G. Leveson. An experimental evaluation of the assumption of independence in multiversion programming. *IEEE Transactions on Software Engineering*, SE-12(1):96–109, 1986. doi: 10.1109/TSE.1986.6312924.

Arinbjörn Kolbeinsson and Benedikt Kolbeinsson. Multimodal alignment for synthetic clinical time series. In *EurIPS 2025 Workshop on Multimodal Representation Learning for Healthcare (MMRL4H)*, Copenhagen, Denmark, 2025. URL https://multimodal-rep-learning-for-health.github.io/papers/12_Multimodal_Alignment_for_Sy.pdf.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, 2023. doi: 10.1145/3600006.3613165.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1612.01474.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021. doi: 10.18653/v1/2021.emnlp-main.565. URL https://arxiv.org/abs/2109.05052.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI Scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Ziming Luo, Atoosa Kasirzadeh, and Nihar B. Shah. The more you automate, the less you see: Hidden pitfalls of AI scientist systems. *arXiv preprint arXiv:2509.08713*, 2025.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, 2024. doi: 10.1038/s42256-024-00832-8.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. DiscoveryBench: Towards data-driven discovery with large language models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2407.01725.

Ludovico Mitchener, Jon M. Laurent, Alex Andonian, Benjamin Tenmann, Siddharth Narayanan, Geemi P. Wellawatte, Andrew White, Lorenzo Sani, and Samuel G. Rodriques. BixBench: a comprehensive benchmark for LLM-based agents in computational biology. *arXiv preprint arXiv:2503.00096*, 2025.

Ileana Montoya Perez, Parisa Movahedi, Valtteri Nieminen, Antti Airola, and Tapio Pahikkala. Does differentially private synthetic data lead to synthetic discoveries? *Methods of Information in Medicine*, 63(1–2):35–51, 2024. doi: 10.1055/a-2385-1355.

NVIDIA. Nemotron 3 Super: Open, efficient mixture-of-experts hybrid Mamba-Transformer model for agentic reasoning. *arXiv preprint arXiv:2604.12374*, 2026.

Joel S. Parker, Michael Mullins, Maggie C. U. Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, John F. Quackenbush, Inge J. Stijleman, Juan Palazzo, J. S. Marron, Andrew B. Nobel, Elaine Mardis, Torsten O. Nielsen, Matthew J. Ellis, Charles M. Perou, and Philip S. Bernard. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, 27(8):1160–1167, 2009. doi: 10.1200/JCO.2008.18.1370.

Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410, 2016. doi: 10.1109/DSAA.2016.49.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all. `https://qwen.ai/blog?id=qwen3.6-35b-a3b`, April 2026.

T. E. Raghunathan, Jerome P. Reiter, and Donald B. Rubin. Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19:1–16, 2003.

Anat Reiner Benaim, Ronit Almog, Yuri Gorelik, Irit Hochberg, Laila Nassar, Tanya Mashiach, Mogher Khamaisi, Yael Lurie, Zaher S. Azzam, Johad Khoury, Daniel Kurnik, and Rafael Beyar. Analyzing medical research results based on synthetic data and their relation to real data results: Systematic comparison from five observational studies. *JMIR Medical Informatics*, 8(2):e16492, 2020. doi: 10.2196/16492.

Zachary T. Rewolinski, Austin V. Zane, Hao Huang, Chandan Singh, Chenglong Wang, Jianfeng Gao, and Bin Yu. Sanity checks for agentic data science. *arXiv preprint arXiv:2604.11003*, 2026.

Lucas Rosenblatt, Bernease Herman, Anastasia Holovenko, Wonkwon Lee, Joshua Loftus, Elizabeth McKinnie, Taras Rumezhak, Andrii Stadnik, Bill Howe, and Julia Stoyanovich. Epistemic parity: Reproducibility as an evaluation metric for differential privacy. *Proceedings of the VLDB Endowment*, 16(11):3178–3191, 2023. doi: 10.14778/3611479.3611517.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL `https://arxiv.org/abs/2310.13548`.

Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, and James Zou. The virtual lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*, 646(8085):716–723, 2025. doi: 10.1038/s41586-025-09442-9.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, 2023. doi: 10.52202/075280-3275. URL `https://arxiv.org/abs/2305.04388`.

Boris van Breugel, Zhaozhi Qian, and Mihaela van der Schaar. Synthetic data, real errors: How (not) to publish and use synthetic data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 34793–34808, 2023.

Xinzhi Wang and Chunwei Liu. The table says otherwise: Testing LLMs with counterfactual relational data. *arXiv preprint arXiv:2606.23667*, 2026.

Kevin Wu, Eric Wu, and James Zou. ClashEval: Quantifying the tug-of-war between an LLM's internal prior and external evidence. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2024. doi: 10.52202/079017-1053. URL `https://arxiv.org/abs/2404.10198`.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations (ICLR)*, 2024. URL `https://arxiv.org/abs/2305.13300`.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional GAN. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Zonglin Yang, Xingtong Liu, and Xinyan Xu. SciIntegrity-Bench: A benchmark for evaluating academic integrity in AI scientist systems. *arXiv preprint arXiv:2605.10246*, 2026.

Jinsung Yoon, Lydia N. Drumright, and Mihaela van der Schaar. Anonymization through data synthesis using generative adversarial networks (ADS-GAN). *IEEE Journal of Biomedical and Health Informatics*, 24(8):2378–2388, 2020. doi: 10.1109/JBHI.2020.2980262.

A Related work

Do conclusions survive on synthetic data? Rosenblatt et al. [2023] propose epistemic parity: whether the conclusions of published studies are reproduced when their analyses are rerun on differentially private synthetic data. Reiner Benaim et al. [2020] and El Emam et al. [2024] replicate clinical analyses on synthetic versions of health datasets; the latter use several synthetic datasets from two generators and find that CTGAN replicates worse. In each case the analysis is fixed in advance and the flaws are whatever the generator produced. Naive inference on synthetic data is also known to inflate the type I error [Decruyenaere et al., 2024] and to yield false discoveries [Montoya Perez et al., 2024]. We differ in three ways: flaws are planted at known places and graded in size, the analysis is chosen by an agent, and each claim is scored by program against the real cohort. This lets us ask at what size a fabricated signal enters an agent's conclusions; the answer depends on the strength of the true markers against which the agent ranks it.

Agents for data-driven discovery. Benchmarks for data-analysis agents score their answers against reference solutions or with an LLM judge [Majumder et al., 2025, Gu et al., 2024, Mitchener et al., 2025]. DiscoveryBench already plants hypotheses in synthetic data (DB-Synth) and asks whether the agent recovers them, scored by a GPT-4 judge. Closest to our work, Rewolinski et al. [2026] propose sanity checks for agentic data science: synthetic outcomes with controlled signal strength, null and perturbed data, and repeated runs of one closed model, GPT-5.2-Codex; they find its confidence poorly calibrated to the stability of its conclusions. We share the idea of planting signals of known strength. We differ in what is planted and what is scored: each run returns a ranked list of claims, each scored by program; we use three open-weight models; one fabrication contradicts textbook biology; the data are a molecular cancer cohort; and we separate whether the agent notices a conflict from whether it rejects the finding. A separate line of work documents agents that fabricate data, leak information or misuse metrics [Luo et al., 2025, Yang et al., 2026]. The failure we measure is different: the agent's analysis is faithful to its data, and the fabrication is in the data.

Prior knowledge against context. Language models can override the context with memorised knowledge [Longpre et al., 2021], accept coherent counter-evidence while favouring their priors when evidence is mixed [Xie et al., 2024], and adopt perturbed retrieved content, subtle errors more readily than unrealistic ones [Wu et al., 2024]. Wang and Liu [2026] test such conflicts on counterfactual relational tables. Separately, the explanations models give need not reflect what drove their answers [Turpin et al., 2023], and models defer to their users' beliefs [Sharma et al., 2024]. Our ESR1 experiment places the conflict in multi-step agentic analysis with a biological prior, and scores both the reported conclusion and its justification: the agent notices that the data contradict the textbook and then explains the result away as a property of the cohort. Deference here is to the data rather than to a user.

Ensembles and common-mode failure. Disagreement within an ensemble is a standard uncertainty signal [Lakshminarayanan et al., 2017], and Deep Generative Ensembles apply it to models trained on synthetic data [van Breugel et al., 2023]; inference pooled across several synthetic datasets goes back to multiple imputation for disclosure limitation [Raghunathan et al., 2003]. We use disagreement for a narrower purpose: a trust score for each claim an agent makes, and a rule for routing claims to real-data validation. Its failure when every generator shares a flaw is expected, since independently built versions of a program share failures [Knight and Leveson, 1986] and errors are correlated across language models [Kim et al., 2025]; our result confirms this limit for agent conclusions and shows how complete it is.

B Implementation details

Cohort. Primary tumours (sample type 01) from the UCSC Xena TCGA breast cancer hub with a PAM50 call, one per patient: 844 tumours (LumA 421, LumB 192, Basal 141, Her2 67, Normal-like 23). The panel holds the 50 PAM50 genes, 19 further well-characterised breast cancer genes (e.g. GATA3, XBP1, TFF1, STARD3, PGAP3, TOP2A, CD274) and 150 little-studied genes: well-expressed genes (mean $\log_2$ expression > 6) whose symbols follow uninformative naming patterns (C#orf#, FAM#, KIAA#, CCDC#, ZNF#, ANKRD#, TMEM#, LRRC#), taking the 150 most variable. Clinical columns: receptor status, HER2 by IHC, age, histology, stage. Mutation flags for TP53,

PIK3CA, CDH1, GATA3 and MAP3K1 come from the MC3 gene-level calls and are missing for tumours outside the MC3 set (30%).

World models. The copula is fitted within each PAM50 subtype: each column is mapped to normal scores by ranks among observed values, the rank correlation is estimated from pairwise-complete observations, shrunk by 5% towards the identity and projected to the nearest positive-definite correlation. Samples are drawn from the multivariate normal and mapped back through each subtype's empirical quantiles; categorical columns take observed values only. Missingness of clinical and mutation columns is copied row-wise from a random real tumour of the same subtype. TVAE and CTGAN are the `ctgan` package implementations trained for 300 epochs with default settings; categorical and mutation columns are modelled as discrete, with missing values as their own category. Each world used for scoring is a fixed 30,000-row sample. The reversal and fabrication edits are applied after sampling, with parameters fixed from that sample and applied identically to the agent's sample and to the world. The bootstrap ensemble members are the copula refitted on bootstrap resamples of the real cohort.

Agent harness. The agent receives a system prompt describing the Python tool and the required output, the research brief, and a working directory holding `cohort.csv` (844 rows) and a data dictionary. Each code block runs in a fresh Python process (pandas, NumPy, SciPy, scikit-learn, statsmodels) with a 180 s limit; output is truncated to 6,000 characters. Reasoning text is removed from the history after each step. Runs stop at the first structured answer or after 20 steps. Sampling follows each model's recommendation: Qwen3.6-35B-A3B (FP8) at temperature 0.6 and top-p 0.95 with thinking, 0.7 and 0.8 without; Nemotron-3-Super-120B-A12B (NVFP4) and Gemma-4-26B-A4B (NVFP4, reasoning returned separately by the server's Gemma reasoning parser) at 1.0 and 0.95, with thinking. All are served with vLLM 0.22.1 on two workstations with one NVIDIA GB10 each. Run seeds set the model's sampling seed and, for generated worlds, the sample drawn; the same seed gives the same first step in every world. The results in the paper come from 408 agent runs.

Scoring. A difference claim is scored as Cohen's d of the variable between the named group and all other rows, an association claim as Spearman's ρ, both with pairwise-complete observations. A claim is supported if its direction holds with $|d| \geq 0.2$ or $|\rho| \geq 0.1$. Claims about categorical variables are unscorable and are counted but not scored. A run flags a variable when a caveat or a sentence of its report names it together with a registered anomaly term (unexpected, contradicts, inconsistent, implausible, anomalous, suspicious, artefact, surprising, opposite, counterintuitive, atypical, conflict, does not match, at odds, questionable); on unedited worlds this never happened for the manipulated genes. Core findings are the claims made in at least half of the real-data runs of a brief. The estimated flaw map marks the 5% of genes with the greatest sum of mean absolute change in correlation with other genes and largest change in subtype difference.

Precision and recall. We use the standard definitions of information retrieval, applied to claims (Figure 2). A relationship is *true* if it holds in the real cohort under the scoring rule above, and *reported* if the agent claims it. A *true positive* (TP) is a reported relationship that is true, a *false positive* (FP) a reported relationship that is false, a *false negative* (FN) a true relationship that is not reported, and a *true negative* (TN) a false relationship that is not reported. Then

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \qquad \text{false-positive rate} = \frac{\text{FP}}{\text{FP} + \text{TN}}.$$

Precision is the share of reported claims that are true, recall the share of true relationships that are reported, and the false-positive rate the share of false relationships that are reported. We compute precision over all scorable claims of a cell and recall over the core findings of a brief (Table 1); for a single true marker such as GRB7, recall is the share of runs that report it. The false relationships an agent could claim are unbounded, so true negatives, and with them the false-positive rate, are defined only for a planted false signal, where the false-positive rate is the share of runs that report it. These claim-level quantities differ from the sample-level precision and recall of fidelity metrics [Alaa et al., 2022], which compare synthetic records with real ones.

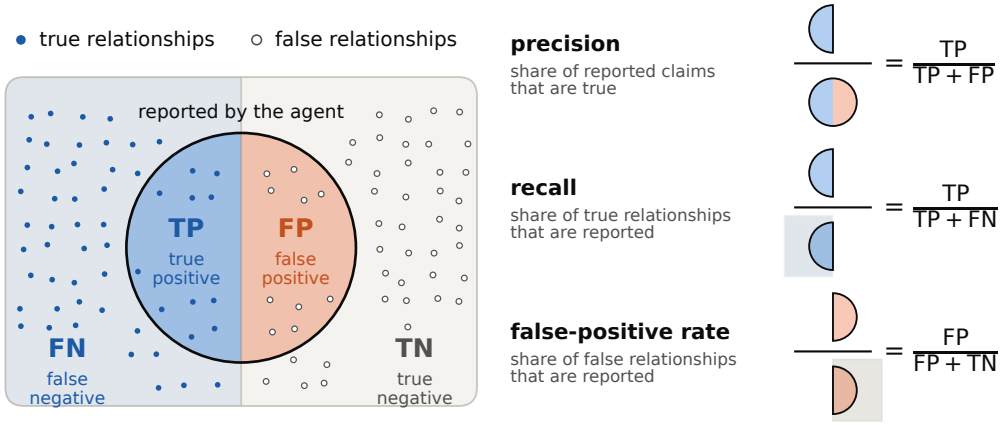


Figure 2: Precision, recall and false-positive rate for an agent's claims. Left: every relationship the agent could claim, filled if true (it holds in the real cohort) and open if false; the disc is what the agent reports. Right: each measure as a fraction of regions of the diagram. A weakened true signal moves findings from TP to FN (lower recall); a fabricated signal adds FP (lower precision and a higher false-positive rate).

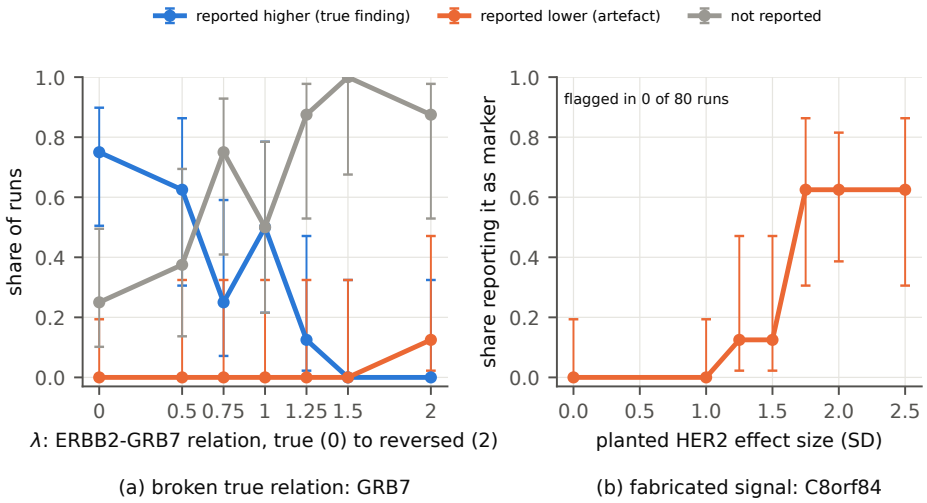


Figure 3: Dose–response on the HER2 targets brief. (a) How the agent reports GRB7 as the reversal edit takes its relation to ERBB2 from true to reversed; Table 2 gives the effect in the synthetic data at each level. (b) Share of runs reporting the planted gene as a HER2 marker. Wilson 95% intervals.

C Additional results

World models. CTGAN trained with default settings collapsed to the marginals: every gene–gene correlation is near zero (ERBB2–GRB7 $r = 0.00$ against 0.84) and there are no subtype differences, although the subtype proportions are right. TVAE is imperfect in familiar ways (attenuated correlations, shrunk variances, HER2-enriched tumours under-sampled at 5% against 8%) but keeps the main subtype differences.

Table 2: Dose–response per cell (HER2 briefs). d: HER2-enriched vs rest effect of the gene in the synthetic world. Higher: runs reporting the gene higher in HER2-enriched tumours, with Wilson 95% interval; flagged: runs naming the gene with an anomaly term.

Brief	Gene	Arm	d	n	Higher	Lower	Omitted	Flagged
neutral	C8orf84	plant_C8orf84_2	+1.68	8	4 [0.22, 0.78]	0	4	0
neutral	C8orf84	copula	-0.11	8	0 [0.00, 0.32]	0	8	0
neutral	GRB7	copula	+1.89	8	6 [0.41, 0.93]	0	2	0
neutral	GRB7	dial_GRB7_2	-0.93	8	0 [0.00, 0.32]	2	6	2
targets	C8orf84	plant_C8orf84_2.5	+2.00	8	5 [0.31, 0.86]	0	3	0
targets	C8orf84	plant_C8orf84_2	+1.68	16	10 [0.39, 0.82]	0	6	0
targets	C8orf84	plant_C8orf84_1.75	+1.50	8	5 [0.31, 0.86]	0	3	0
targets	C8orf84	plant_C8orf84_1.5	+1.30	8	1 [0.02, 0.47]	0	7	0
targets	C8orf84	plant_C8orf84_1.25	+1.09	8	1 [0.02, 0.47]	0	7	0
targets	C8orf84	plant_C8orf84_1	+0.86	16	0 [0.00, 0.19]	0	16	0
targets	C8orf84	copula	-0.11	16	0 [0.00, 0.19]	0	16	0
targets	GRB7	copula	+1.89	16	12 [0.51, 0.90]	0	4	0
targets	GRB7	real	+1.86	8	7 [0.53, 0.98]	0	1	0
targets	GRB7	dial_GRB7_0.5	+1.54	8	5 [0.31, 0.86]	0	3	0
targets	GRB7	dial_GRB7_0.75	+1.19	8	2 [0.07, 0.59]	0	6	0
targets	GRB7	dial_GRB7_1	+0.71	8	4 [0.22, 0.78]	0	4	0
targets	GRB7	dial_GRB7_1.25	+0.18	8	1 [0.02, 0.47]	0	7	0
targets	GRB7	dial_GRB7_1.5	-0.29	8	0 [0.00, 0.32]	0	8	0
targets	GRB7	dial_GRB7_2	-0.93	8	0 [0.00, 0.32]	1	7	3
targets	GRB7	real_boot	+1.86	8	7 [0.53, 0.98]	0	1	0

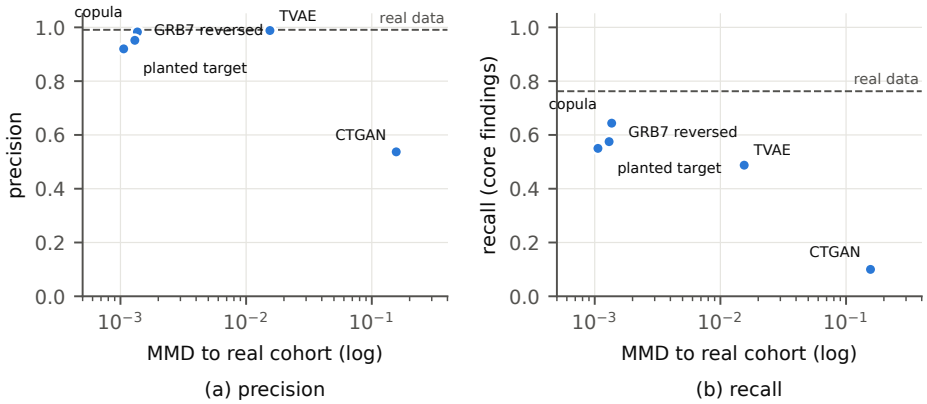


Figure 4: Precision and recall of the agent's conclusions against each world's MMD to the real cohort (HER2 targets brief).

Table 3: Claim and run dependence. Claims from one run share a brief, a world and an analysis, and the same claim recurs across runs, so the intervals here resample whole runs within each cell (2,000 replicates). The table gives the precision and recall of Table 1 with 95% intervals (eight runs per cell; core findings held fixed). For the trust signal (Section 3.3), the 3,906 claims from 368 runs are 1,089 distinct pairs of world and claim (1,006 correct, 83 artefacts). Counted once per pair, the AUROC is 0.9993 with the copula refits and 0.984 without them; resampling runs within the 40 cells of model, brief and world gives 95% intervals of 0.9996–0.99997 and 0.981–0.987.

	HER2 targets brief		Basal brief	
World	Precision	Recall	Precision	Recall
Real data	0.99 [0.97, 1.00]	0.76 [0.60, 0.89]	1.00 [1.00, 1.00]	0.64 [0.51, 0.76]
Copula	0.96 [0.93, 0.99]	0.59 [0.45, 0.70]	1.00 [1.00, 1.00]	0.54 [0.47, 0.60]
TVAE	0.99 [0.96, 1.00]	0.49 [0.36, 0.61]	1.00 [1.00, 1.00]	0.43 [0.29, 0.54]
CTGAN	0.54 [0.45, 0.66]	0.10 [0.04, 0.16]	0.44 [0.35, 0.52]	0.01 [0.00, 0.03]

Table 4: Precision (share of scorable claims true on the real cohort) and faithfulness (share supported by the sample the agent analysed), HER2 targets brief.

	Qwen3.6, thinking			Qwen3.6, no thinking			Nemotron-120B			Gemma-4-26B		
Arm	n	Prec.	Faith.	n	Prec.	Faith.	n	Prec.	Faith.	n	Prec.	Faith.
real	8	0.99	0.99	–	–	–	8	0.87	0.87	–	–	–
copula	16	0.98	0.99	8	0.96	0.98	8	0.95	0.95	8	1.00	1.00
dial_GRB7_2	8	0.95	1.00	8	0.93	0.97	15	0.89	0.90	8	0.97	1.00
plant_C8orf84_2	16	0.92	0.99	8	0.86	0.93	16	0.85	0.88	8	0.90	0.99
plant_ESR1_3	16	0.91	0.99	7	0.88	1.00	–	–	–	8	0.93	1.00

Table 5: Fabricated signals against and without prior knowledge. ESR1 (textbook lower in HER2-enriched tumours) and FAM84A (little studied) shifted to a synthetic effect of $+1.7$ to $+1.8$ SD. Transmitted: reported, not flagged; rationalised: reported and flagged as unexpected; rejected: flagged and not reported; prior: the textbook direction asserted against the analysed data.

Agent	Gene	Brief	n	Transmitted	Rationalised	Rejected	Prior	Omitted
Qwen3.6, thinking	ESR1	targets	16	1	12	2	0	1
Qwen3.6, thinking	ESR1	neutral	16	7	9	0	0	0
Qwen3.6, no thinking	ESR1	targets	7	1	4	1	0	1
Nemotron-120B	ESR1	targets	8	1	1	0	1	5
Gemma-4-26B	ESR1	targets	8	2	2	0	0	4
Qwen3.6, thinking	FAM84A	targets	15	6	0	0	0	9

Table 6: Replication on the basal-like brief. The top true basal markers have $|d|$ of 2.1–2.4.

Gene	World	n	Higher in basal	Lower	Omitted	Flagged
FOXC1	real	8	7 [0.53, 0.98]	0	1	0
FOXC1	copula	8	5 [0.31, 0.86]	0	3	0
FOXC1	FOXC1 reversed ($\lambda = 2$)	8	0 [0.00, 0.32]	0	8	0
C8orf84	copula	8	0 [0.00, 0.32]	0	8	0
C8orf84	planted, $d = 1.74$	8	1 [0.02, 0.47]	0	7	0
C8orf84	planted, $d = 2.08$	8	5 [0.31, 0.86]	0	3	0

Table 7: Targeted validation. Oracle map: the edited gene. Estimated map: the 5% of genes whose correlation row or subtype-difference profile deviates most from the real cohort's. Prec.: precision before validation; after: precision once routed claims are validated. Routed: share of claims sent to validation; caught: share of failing claims among them.

				Oracle map			Estimated map		
Arm	Claims	Fail	Prec.	Routed	Caught	After	Routed	Caught	After
plant_C8orf84_2	296	18	0.94	0.05	0.78	0.99	0.15	0.78	0.98
plant_ESR1_3	413	30	0.93	0.13	0.97	1.00	0.20	0.97	1.00
plant_FAM84A_2	135	9	0.93	0.05	0.78	0.98	0.11	0.78	0.98
dial_GRB7_2	234	8	0.97	0.02	0.50	0.98	0.12	0.50	0.98
tvae	180	1	0.99	0.00	0.00	0.99	0.08	0.00	0.99
ctgan	139	73	0.47	0.00	0.00	0.47	0.05	0.04	0.47

Table 8: What the agent model knows. Band: share of five answers giving the correct chromosome band (NCBI Gene). Direction: answers to "is *gene* higher, lower or no different in HER2-enriched tumours?". With thinking, some answers began with a preamble and could not be parsed.

			No thinking		Thinking	
Gene	Band	Truth	Band	Direction	Band	Direction
GRB7	17q12	higher	0.0	higher 5	0.4	higher 5 (0 unparsed)
PGAP3	17q12	higher	0.0	higher 5	0.0	higher 4 (1 unparsed)
STARD3	17q12	higher	0.0	higher 5	0.0	higher 5 (0 unparsed)
PSMD3	17q21.1	higher	0.0	higher 5	0.0	higher 2 (3 unparsed)
C17orf37	17q12	higher	0.0	higher 5	0.0	higher 1 (4 unparsed)
C8orf84	8q21.11	none	0.0	higher 3, unknown 2	0.0	unknown 4 (1 unparsed)
ESR1	6q25.1	lower	1.0	lower 5	0.8	lower 4 (1 unparsed)
FOXA1	14q21.1	lower	0.4	lower 4, higher 1	0.0	lower 4 (1 unparsed)
TMEM139	7q34	none	0.0	higher 5	0.0	higher 4 (1 unparsed)
CCDC74A	2q21.1	none	0.0	higher 5	0.0	higher 1, unknown 1 (3 unparsed)