Appendix

This appendix collects figures referenced in the main text but omitted from the body chapters for reasons of space, or because their information is fully captured by the aggregated versions shown there. The first part provides the full per-provider, per-query grids for the embedding-distance versus DL-distance analysis of Section 4.2; the second adds the Manhattan metric omitted from the main figures.

Embedding distance vs. DL distance: per-provider grids. Each provider's data are split across four panels of five queries each, in the order listed in Table 3.1, to keep each individual panel legible; columns are the four distance metrics (chebyshev, cosine, euclidean, manhattan). The colour convention matches the aggregated figures of Section 4.2: red shading indicates that the embedding accumulates displacement faster than the DL budget grows, green the opposite. Both axes are normalised to $[0, 1]$ by the per-query maximum of each curve.

These grids make the within-provider query-level variability directly visible: the cross-query spread seen in the standard-deviation bands of the aggregated figures decomposes here into individual query trajectories. The short-query effect described in Section 4.2, whereby queries such as *"what is a fandango"* (length 18) saturate earlier and show a wider red band than long queries such as *"what can you do with a degree in mathematics besides suffering?"* (length 63), is visible in every provider, with the short-query rows occupying the earlier parts of each provider's sequence.

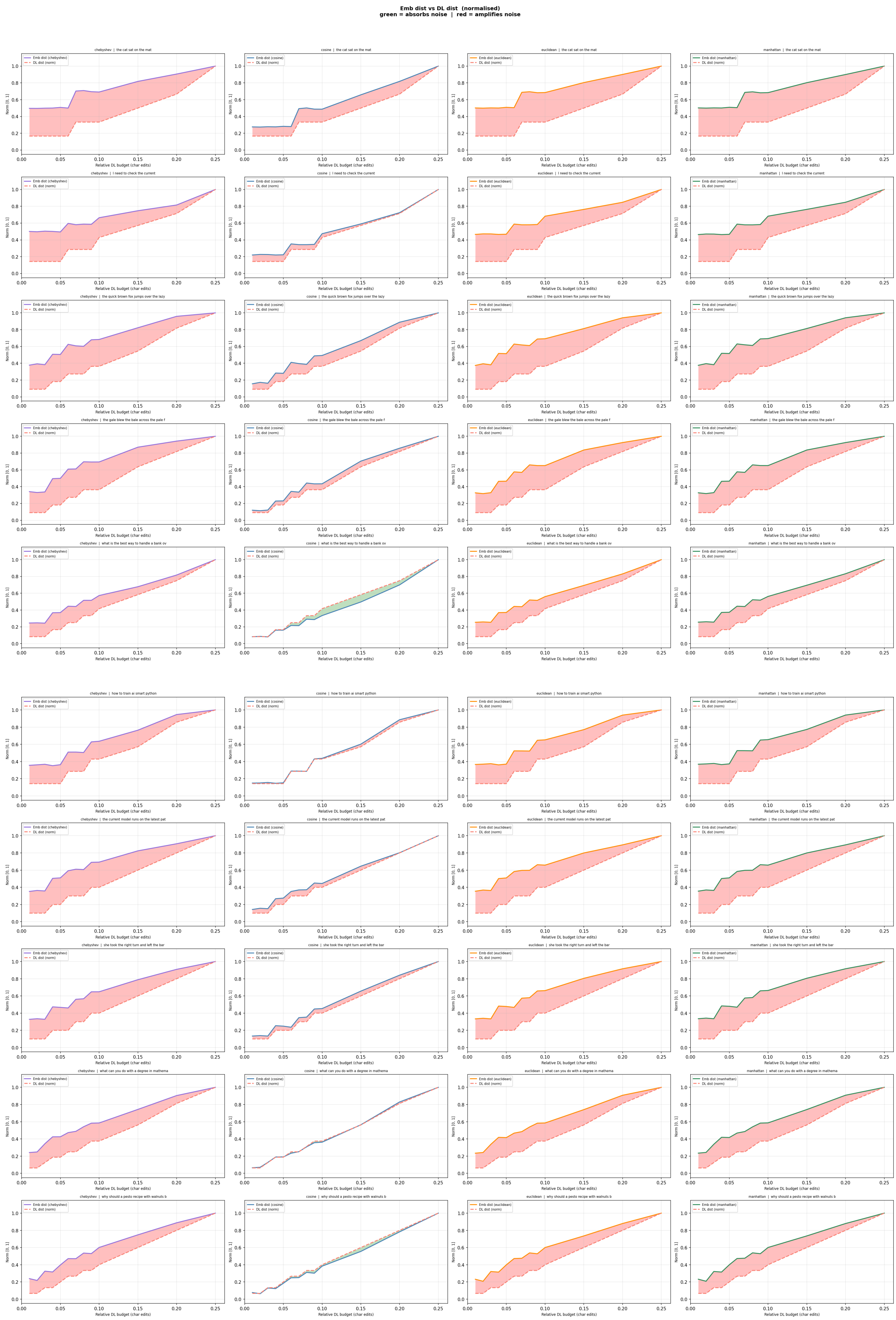

Figure 20. Embedding distance vs. DL distance, **Azure** (`text-embedding-3-large`), queries 1–10. Columns: chebyshev, cosine, euclidean, manhattan.

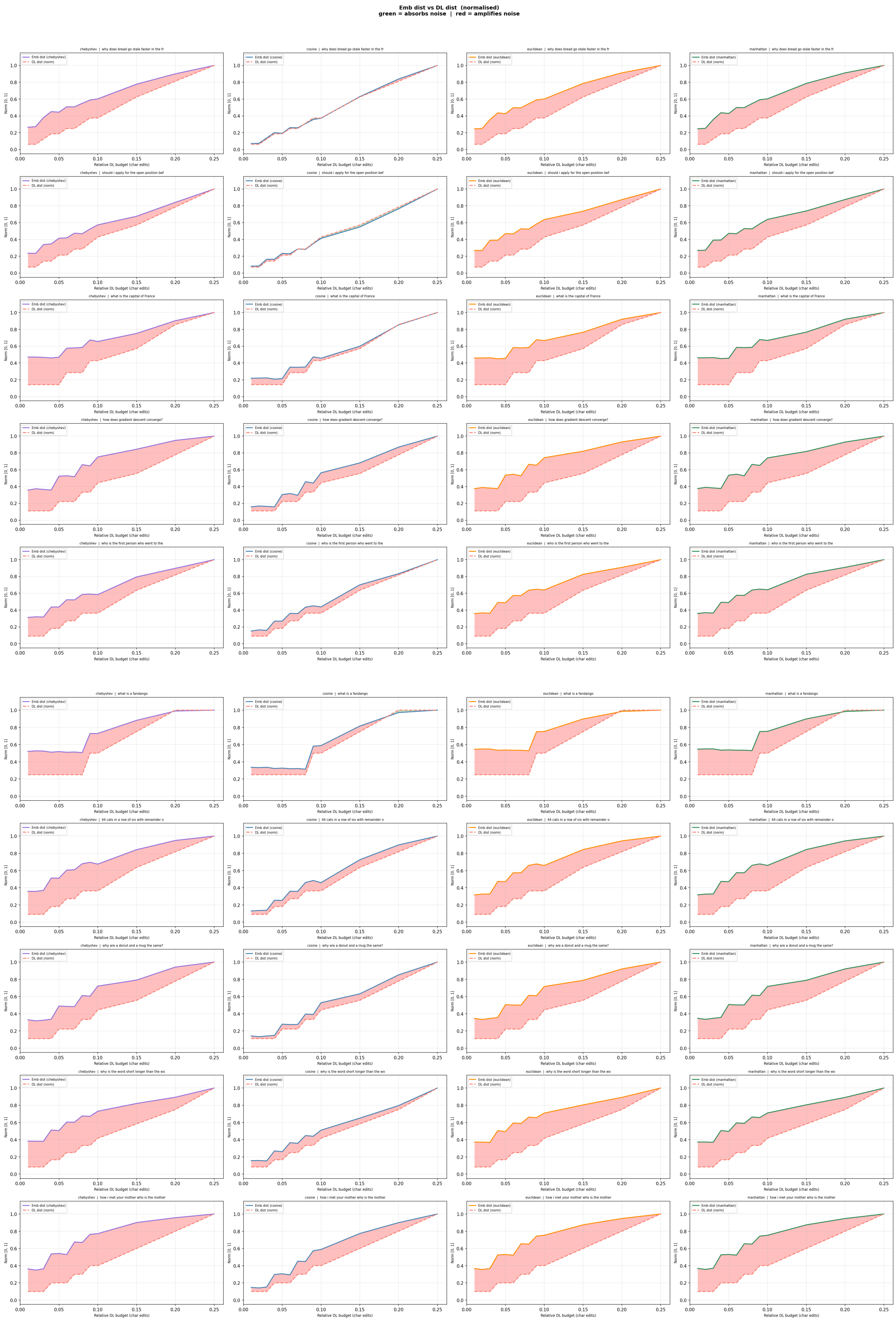

Figure 21. Embedding distance vs. DL distance, **Azure** (`text-embedding-3-large`), queries 11–20.

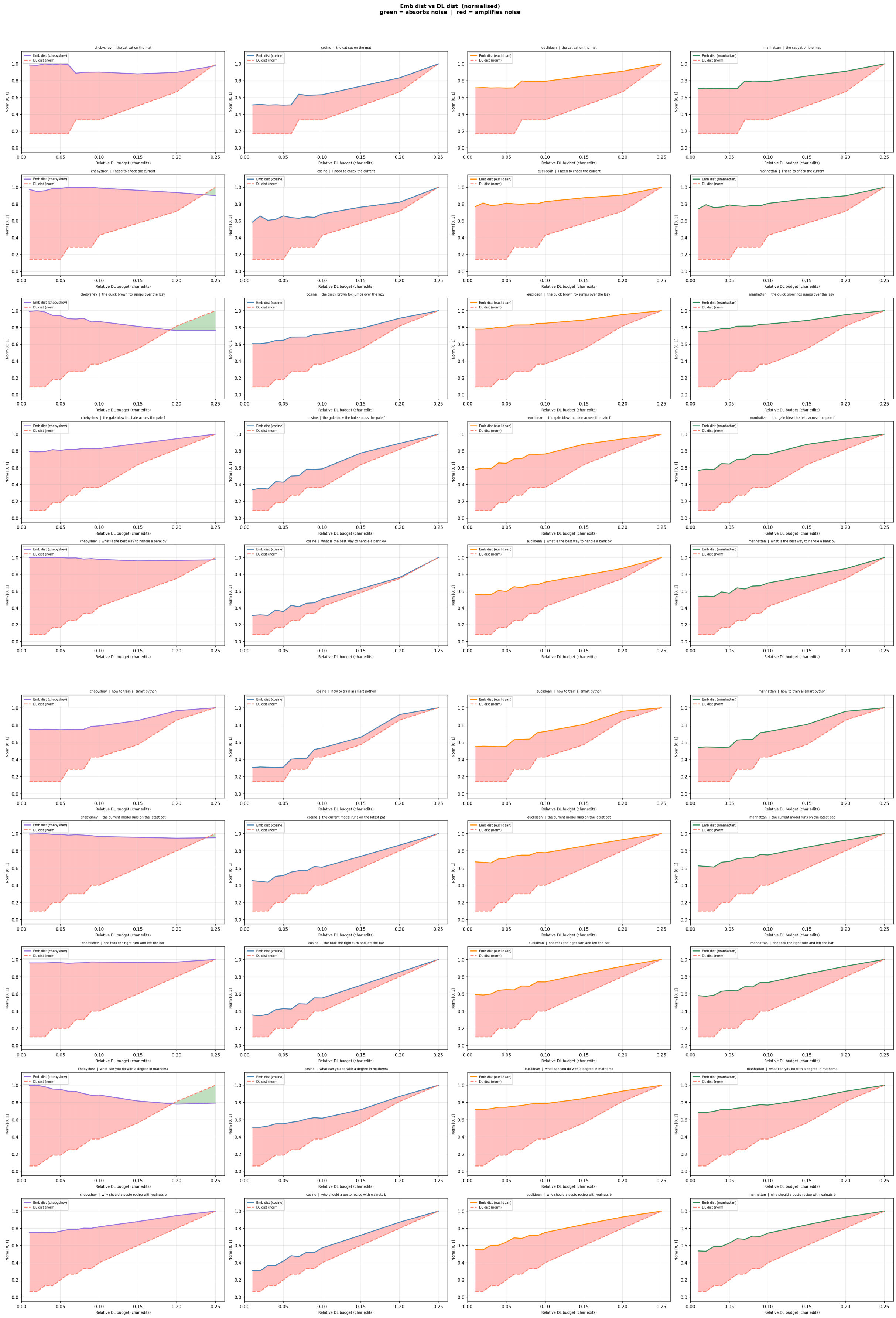

Figure 22. Embedding distance vs. DL distance, **Gemini** (`gemini-embedding-001`), queries 1–10. Data cover only the lower portion of the ρ grid ($\rho \leq 0.15$) owing to API rate constraints. Near-immediate saturation is visible across all query rows and all metrics.

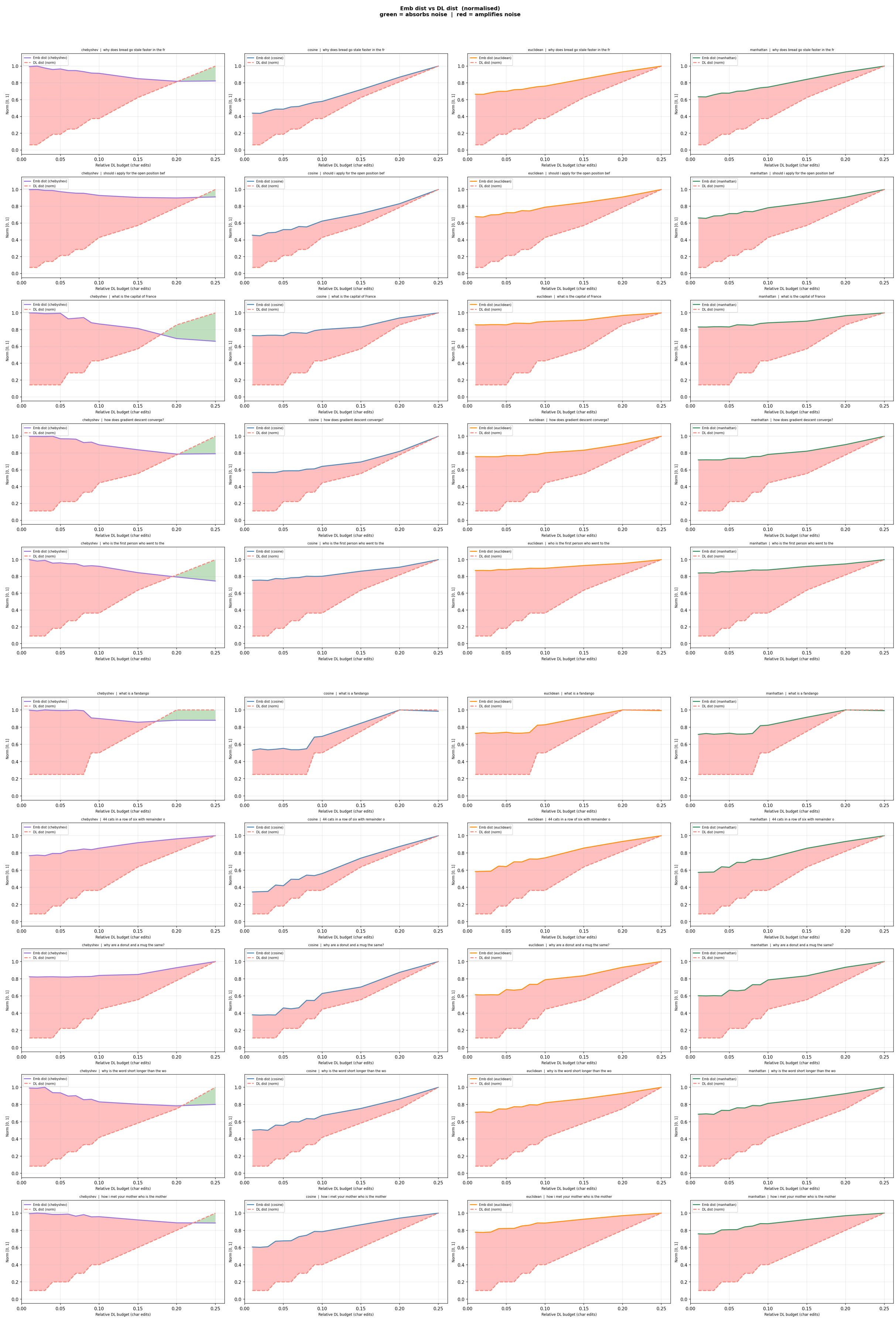

Figure 23. Embedding distance vs. DL distance, **Gemini** (`gemini-embedding-001`), queries 11–20.

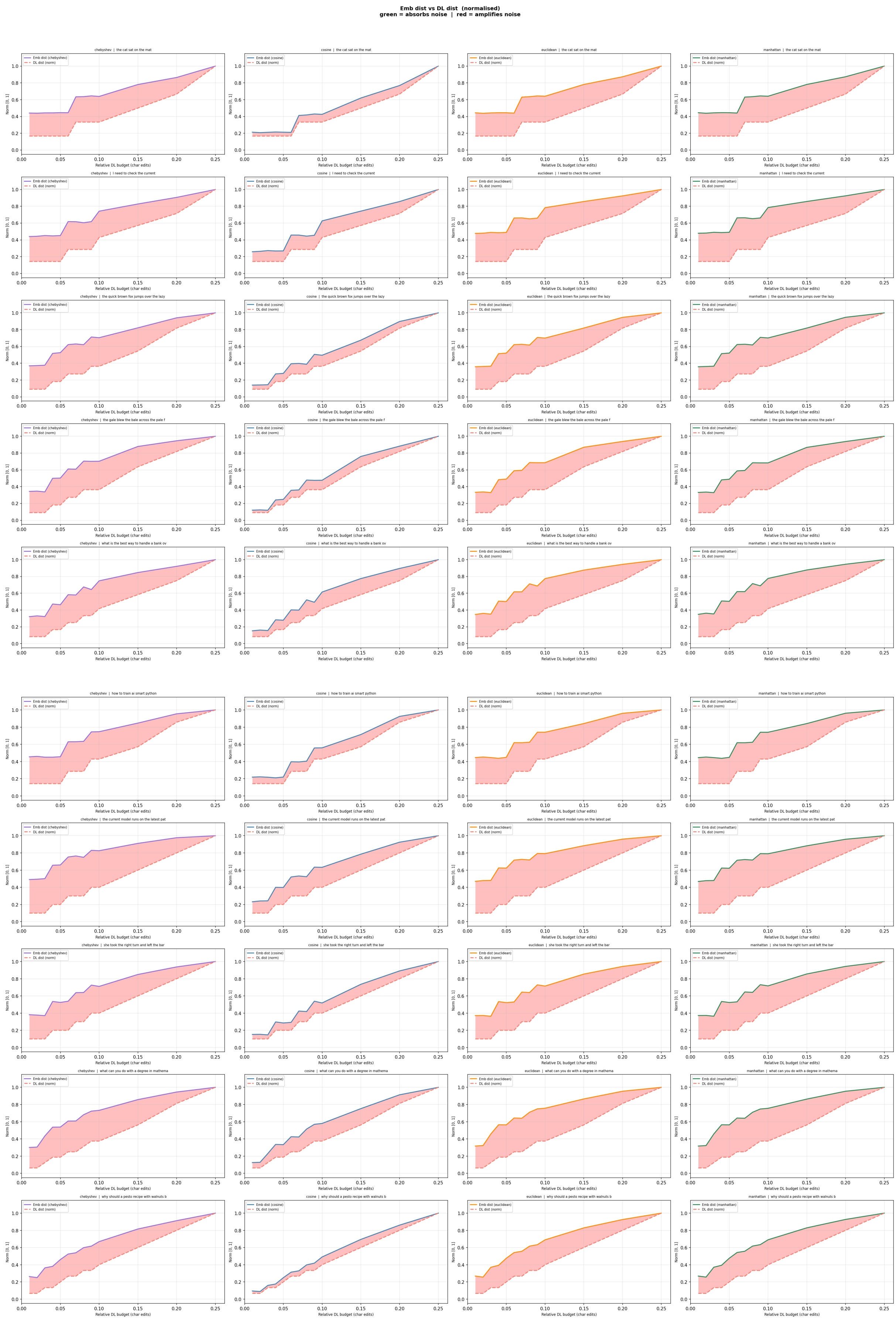

Figure 24. Embedding distance vs. DL distance, **HF_bge** (`BAAI/bge-m3`), queries 1–10. The early-plateau then abrupt-jump pattern described in Section 4.2 is visible in the short-query rows under cosine.

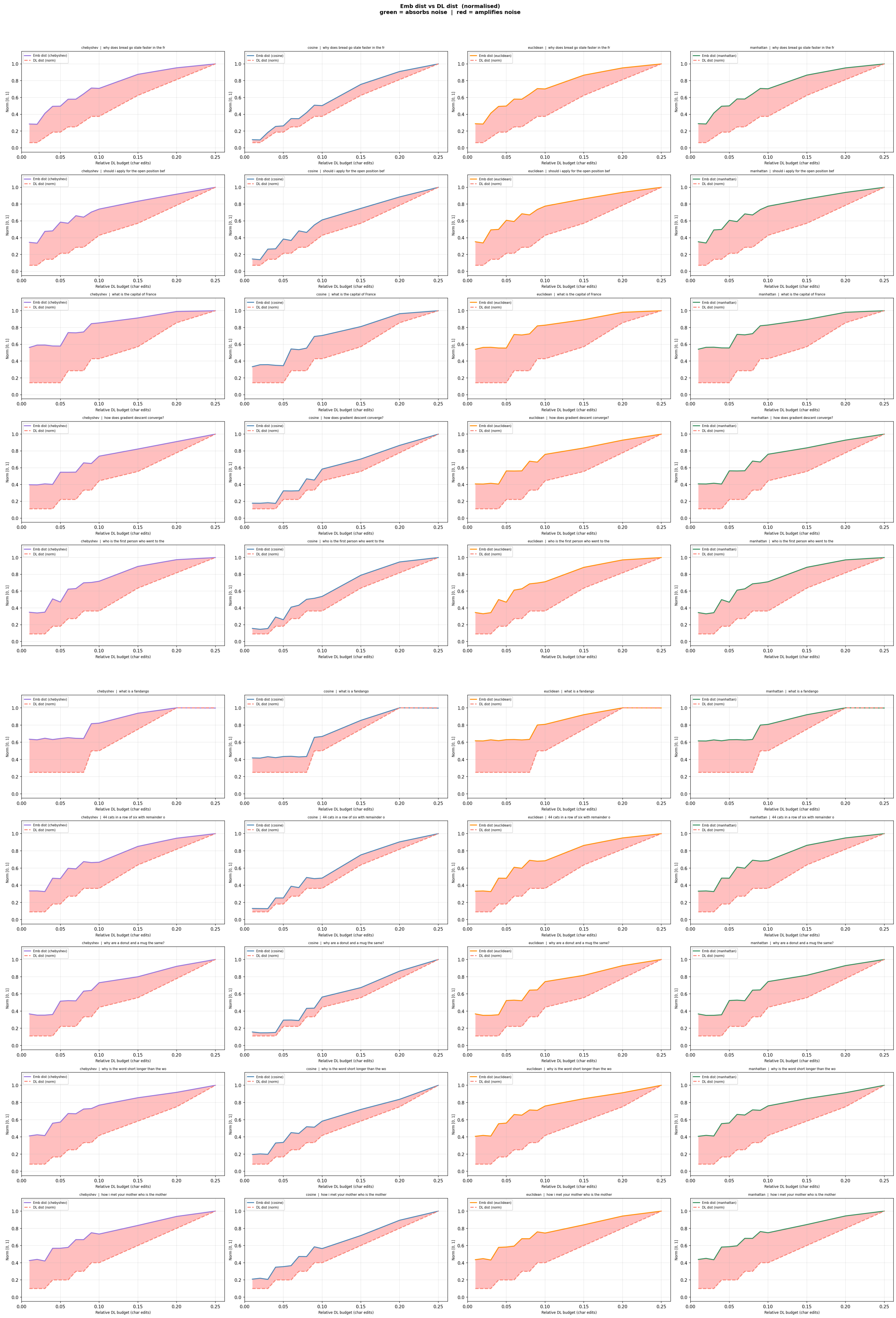

Figure 25. Embedding distance vs. DL distance, **HF_bge** (`BAAI/bge-m3`), queries 11–20.

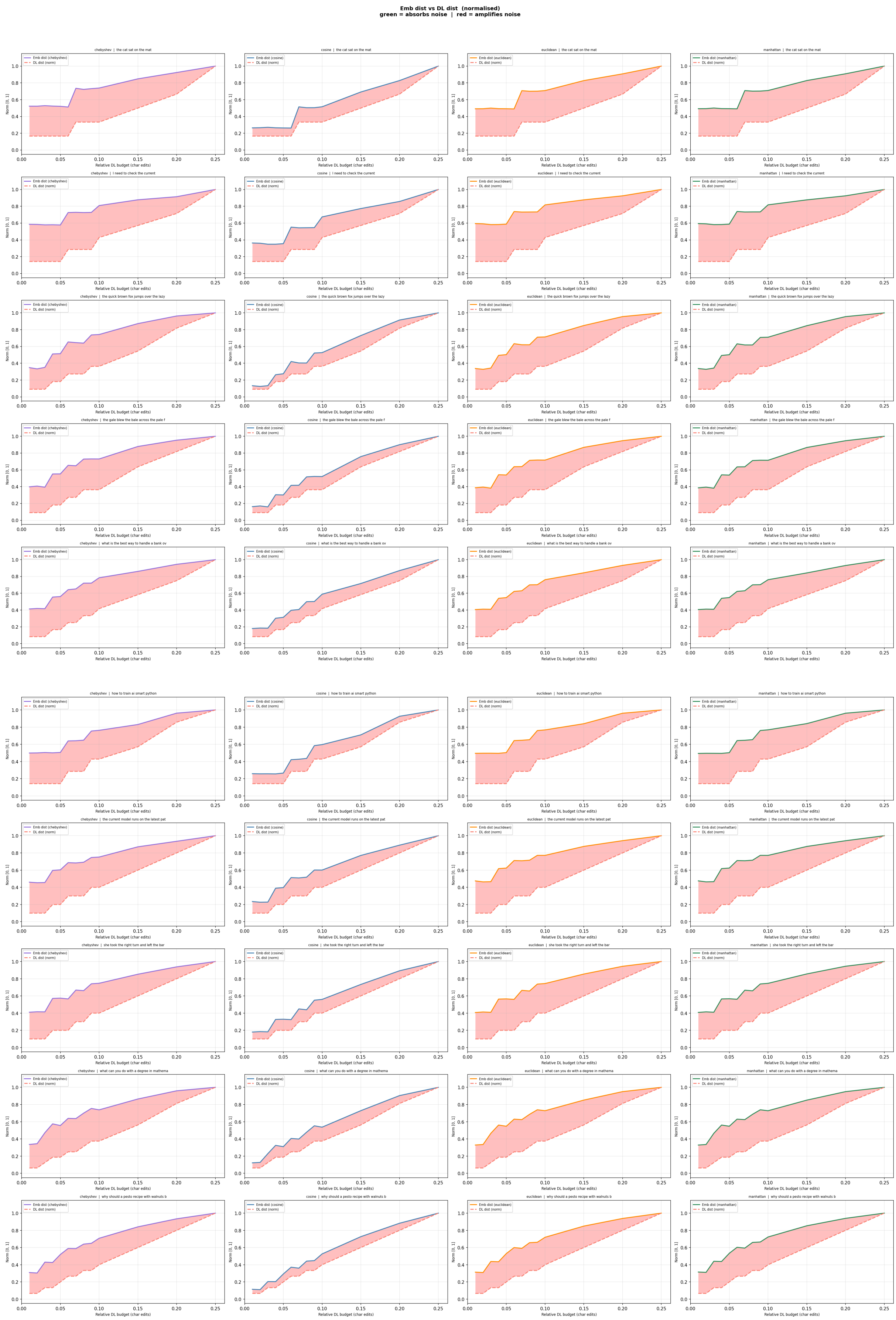

Figure 26. Embedding distance vs. DL distance, **HF_e5** (`intfloat/multilingual-e5-large`), queries 1–10. The cosine column shows the closest tracking of the DL diagonal among all providers, with the narrowest per-query spread and the most persistent green absorption zones at low budgets.

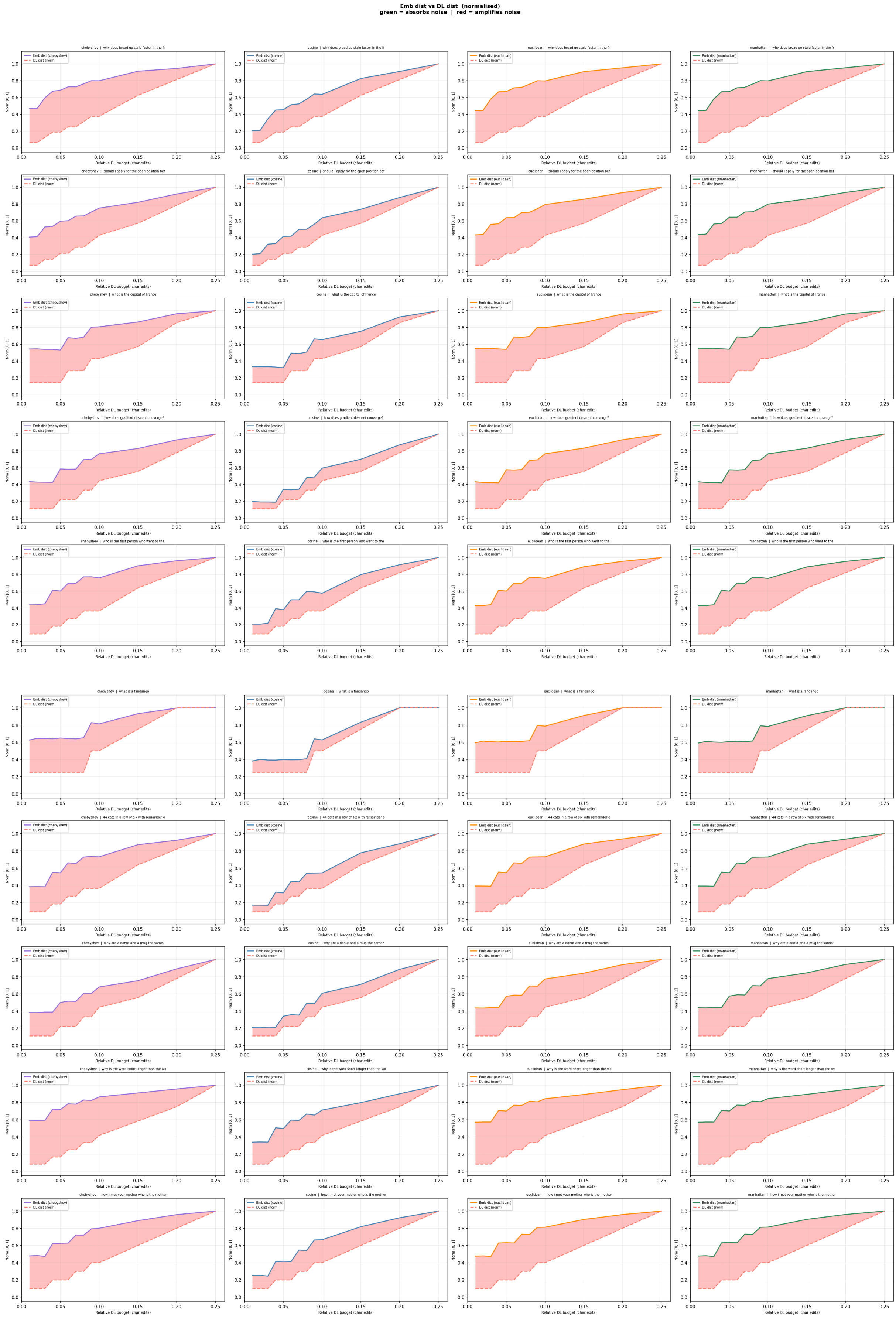

Figure 27. Embedding distance vs. DL distance, **HF_e5** (`intfloat/multilingual-e5-large`), queries 11–20.

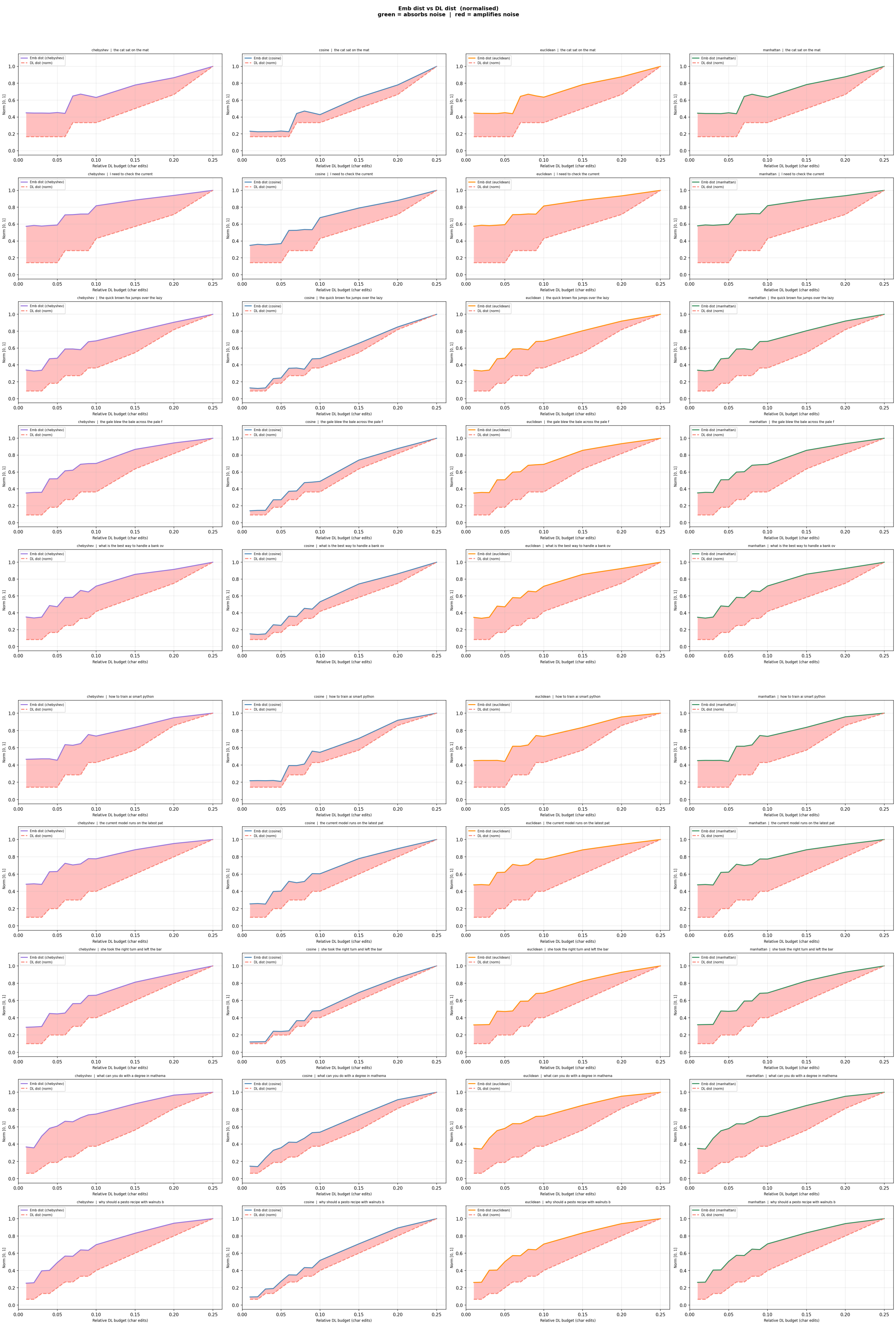

Figure 28. Embedding distance vs. DL distance, **Ollama** (`nomic-embed-text`), queries 1–10. The wide red bands and query-level variability under cosine are consistent with the high distortion index ($\kappa_{\cos} = 169$) reported in Table 4.5.

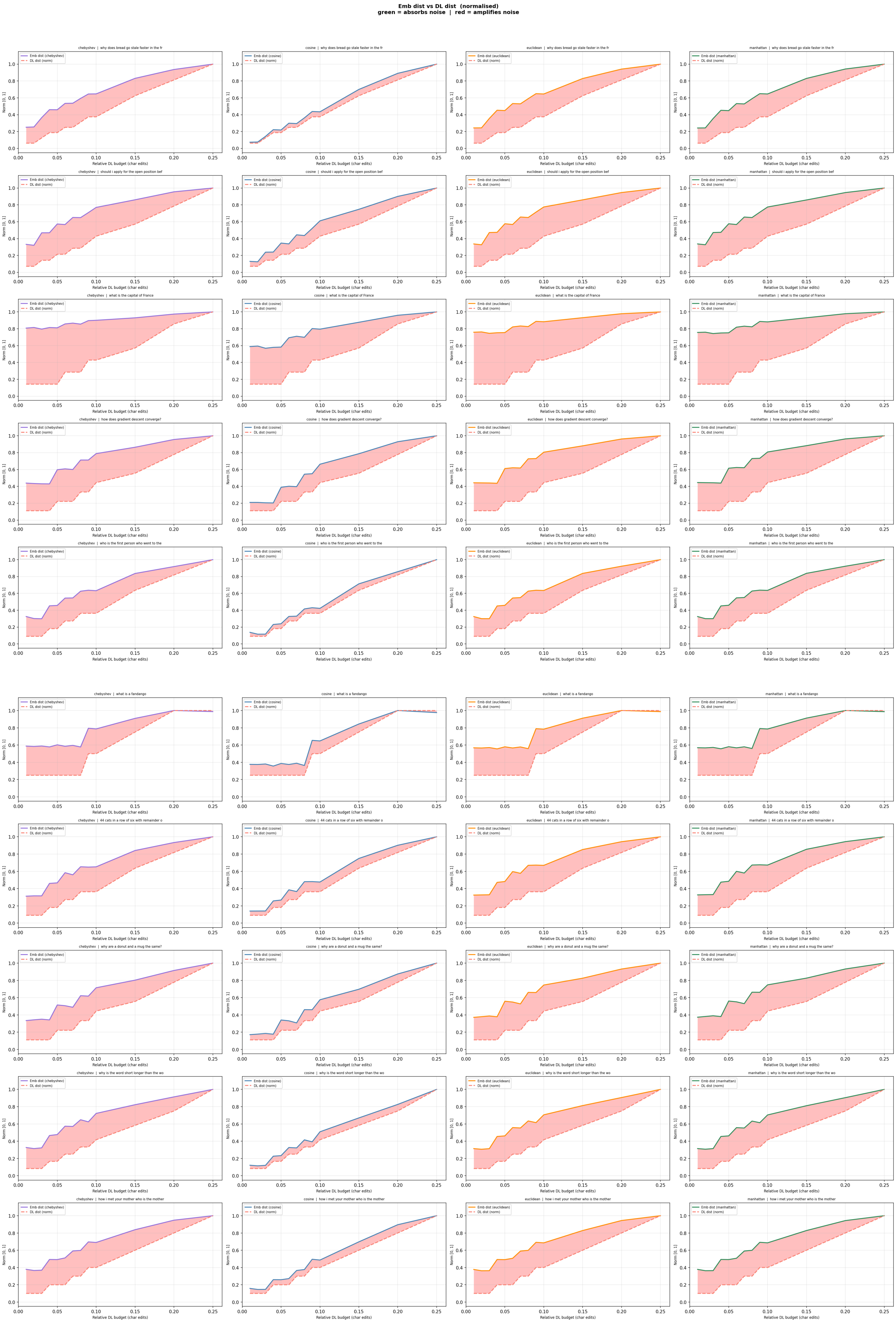

Figure 29. Embedding distance vs. DL distance, **Ollama** (`nomic-embed-text`), queries 11–20.

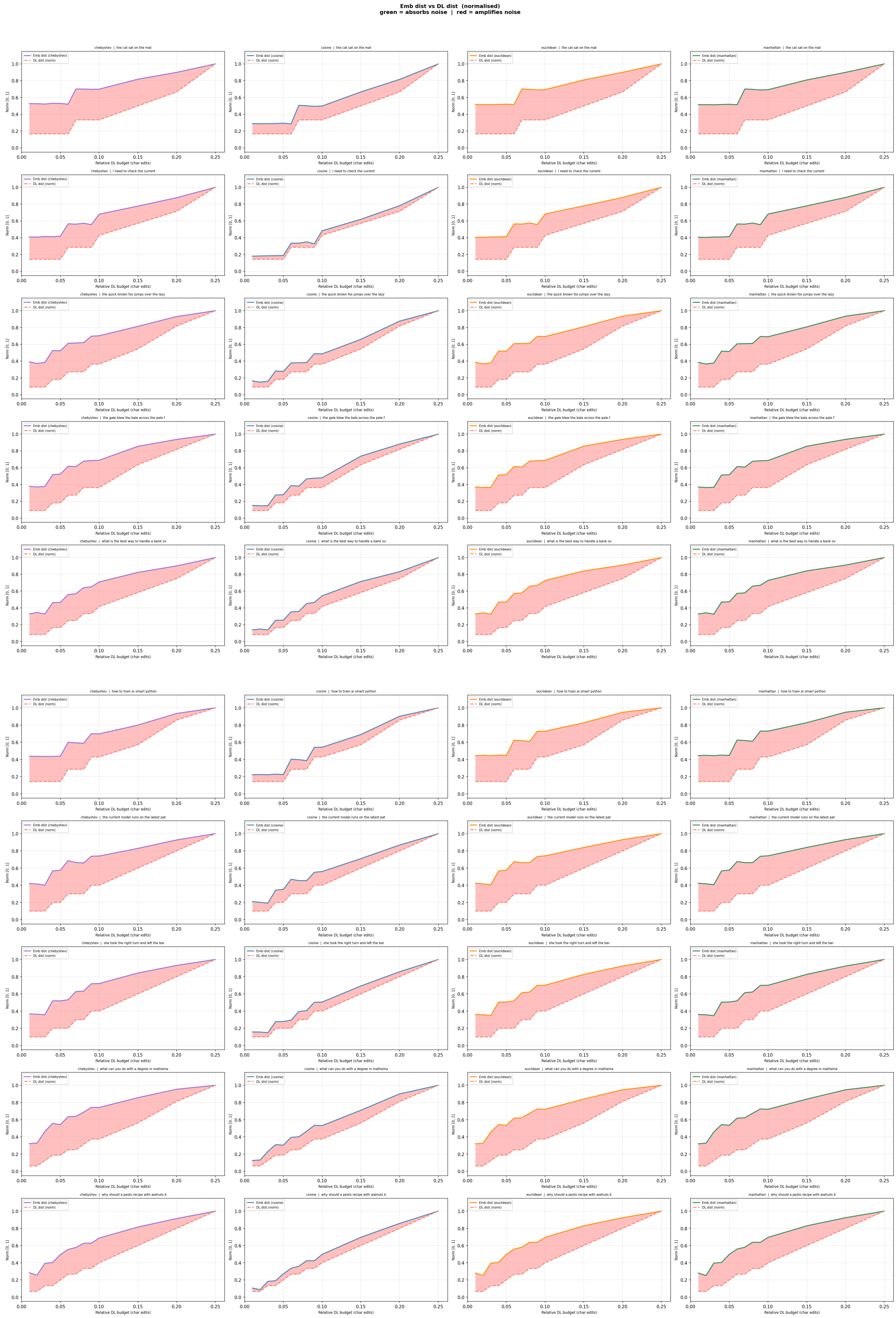

Figure 30. Embedding distance vs. DL distance, **OR_nvidia** (`nvidia/llama-nemotron-embed-vl-1b-v2`), queries 1–10. Behaviour is intermediate between `HF_e5` and the higher-distortion providers; the cosine column shows moderate front-loading consistent with $\bar{\kappa}_{\cos} = 12.95$ at $\rho^* = 0.08$.

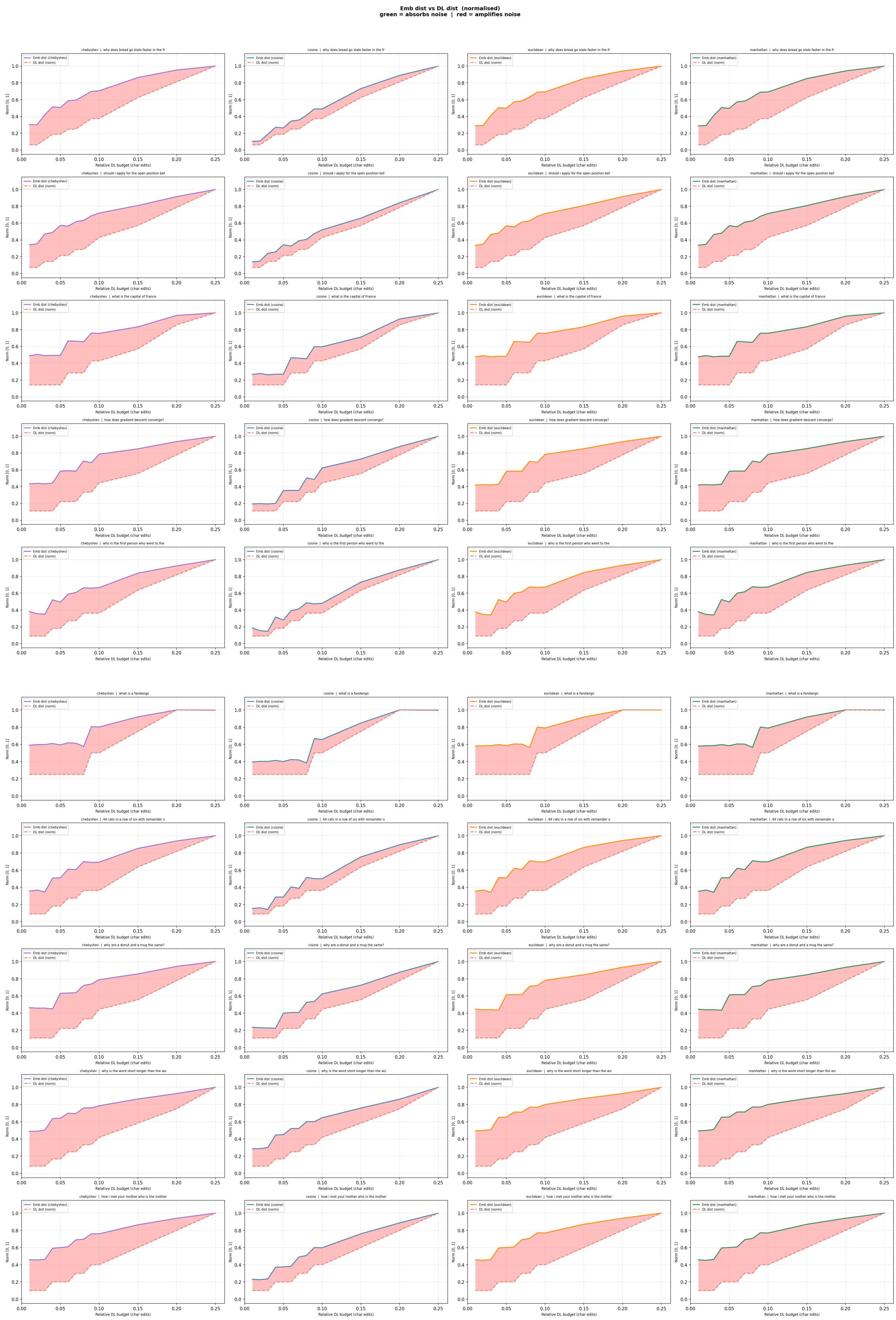

Figure 31. Embedding distance vs. DL distance, **OR_nvidia** (`nvidia/llama-nemotron-embed-v1-1b-v2`), queries 11–20.

Embedding distance vs. DL distance: Manhattan metric. The following figure completes the metric coverage for the aggregated embedding-distance analysis. As noted in Section 4.2, the Manhattan (ℓ_1) metric produces patterns nearly identical to the Euclidean (ℓ_2) result (Figure 4.6) for every provider, with mean curves and standard-deviation bands almost superimposed. It is included here for completeness.

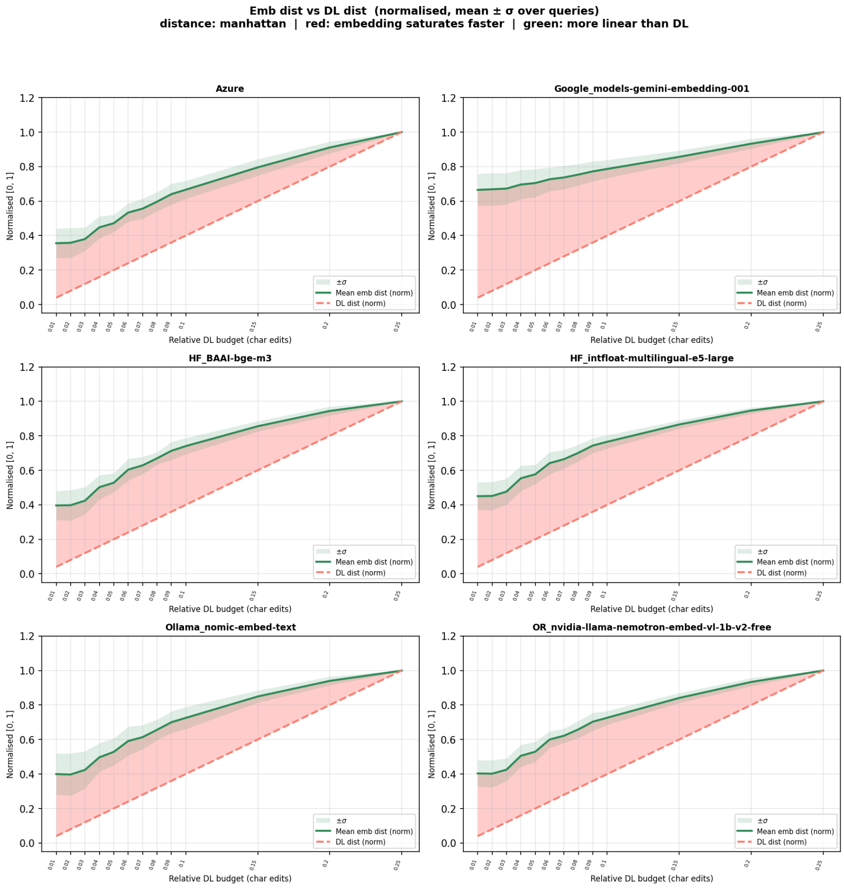

Figure 32. Mean normalised embedding distance vs. normalised DL distance under the **Manhattan** (ℓ_1) metric. Layout and colour convention as in Figure 4.5. The pattern is near-identical to the Euclidean result (Figure 4.6).

References

[1] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[2] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423/.

[3] Tom Brown et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[4] Patrick Lewis et al. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. 2020.

[5] Lorenzo Canale et al. "BES4RAG: A Framework for Embedding Model Selection in Retrieval-Augmented Generation". In: *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*. Ed. by Cristina Bosco et al. Cagliari, Italy: CEUR Workshop Proceedings, Sept. 2025, pp. 134–142. ISBN: 979-12-243-0587-3. URL: https://aclanthology.org/2025.clicit-1.15/.

[6] Liang Wang et al. "Improving Text Embeddings with Large Language Models". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. 2024, pp. 11897–11916. URL: https://aclanthology.org/2024.acl-long.642.

[7] Yougang Guo et al. "Enhancing Noise Robustness of Retrieval-Augmented Language Models with Adaptive Adversarial Training". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. 2024. URL: https://aclanthology.org/2024.acl-long.540/.

[8] Jabez Magomere et al. "When Claims Evolve: Evaluating and Enhancing the Robustness of Embedding Models Against Misinformation Edits". In: *Findings of the Association for Computational Linguistics: ACL 2025*. Vienna, Austria:

Association for Computational Linguistics, 2025, pp. 22374–22404. DOI: 10.18653/v1/2025.findings-acl.1150.

[9] Tanmay Chavan et al. "SenTest: Evaluating Robustness of Sentence Encoders". In: *CoRR* abs/2311.17722 (2023). DOI: 10.48550/ARXIV.2311.17722. arXiv: 2311.17722. URL: https://doi.org/10.48550/arXiv.2311.17722.

[10] Reagan J. Lee, Samarth Goel, and Kannan Ramchandran. *Quantifying Positional Biases in Text Embedding Models*. 2025. arXiv: 2412.15241 [cs.CL]. URL: https://arxiv.org/abs/2412.15241.

[11] Andrés Marzal and Sergio Barrachina. *Edit Distance Embedding using Convolutional Neural Networks*. 2020. arXiv: 2001.11692 [cs.LG]. URL: https://arxiv.org/abs/2001.11692.

[12] E. Kreyszig. *Introductory Functional Analysis with Applications*. Wiley Classics Library. Wiley, 1991. ISBN: 9780471504597. URL: https://books.google.it/books?id=AQtMEAAAQBAJ.

[13] W. Rudin. *Functional Analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, 1991. ISBN: 9780070542365. URL: https://books.google.it/books?id=Sh_vAAAAMAAJ.

[14] Manfredo Perdigão do Carmo. *Riemannian Geometry*. Mathematics: Theory and Applications. Boston: Birkhäuser, 1992. ISBN: 0817634908.

[15] Gonzalo Navarro. "A Guided Tour to Approximate String Matching". In: *ACM Computing Surveys* 33 (Apr. 2000). DOI: 10.1145/375360.375365.

[16] Fred J. Damerau. "A Technique for Computer Detection and Correction of Spelling Errors". In: *Commun. ACM* 7.3 (Mar. 1964), pp. 171–176. ISSN: 0001-0782. DOI: 10.1145/363958.363994. URL: https://doi.org/10.1145/363958.363994.

[17] Gregory V. Bard. "Spelling-Error Tolerant, Order-Independent Pass-Phrases via the Damerau-Levenshtein String-Edit Distance Metric". In: *Proceedings of the Fifth Australasian Symposium on ACSW Frontiers*. ACSW '07. AUS: Australian Computer Society, Inc., 2007, pp. 117–124.

[18] J. R. Norris. *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997.

[19] S. Karlin and H. E. Taylor. *A First Course in Stochastic Processes*. Academic Press, 2012. ISBN: 9780080570419. URL: https://books.google.it/books?id=dSDxjX9nmmMC.

[20] Franco Fagnola and Emanuela Sasso. "Catene di Markov". Lecture notes, Politecnico di Milano. Sept. 2017. URL: https://www.mate.polimi.it.

[21] M. Reed and B. Simon. *Methods of Modern Mathematical Physics: Functional Analysis*. Methods of Modern Mathematical Physics v. 1. Academic Press, 1980. ISBN: 9780125850506. URL: https://books.google.it/books?id=hInvAAAAMAAJ.

[22] Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. "On the Surprising Behavior of Distance Metrics in High Dimensional Spaces". In: *Proceedings of the 8th International Conference on Database Theory*. ICDT '01. Berlin, Heidelberg: Springer-Verlag, 2001, pp. 420–434. ISBN: 3540414568.

[23] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL: http://www.deeplearningbook.org.

[24] Saul B. Needleman and Christian D. Wunsch. "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins". In: *Journal of Molecular Biology* 48.3 (1970), pp. 443–453. ISSN: 0022-2836. DOI: https://doi.org/10.1016/0022-2836(70)90057-4. URL: https://www.sciencedirect.com/science/article/pii/0022283670900574.

[25] Tamás Biró. *Minimal Edit Distance is a Distance Metric — Remarks on Assignment 2, LING 227*. 2014. URL: http://birot.hu/courses/2014-LC/MED-is-distance.pdf.

[26] David Williams. *Probability with Martingales*. Cambridge Mathematical Textbooks. Cambridge University Press, 1991. ISBN: 978-0-521-40605-5.

[27] E. Seneta. *Non-Negative Matrices and Markov Chains*. Springer Series in Statistics. New York, NY: Springer, 2006. DOI: 10.1007/0-387-32792-4. URL: https://cds.cern.ch/record/942627.

[28] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990. ISBN: 0521386322.

[29] Rico Sennrich, Barry Haddow, and Alexandra Birch. "Neural Machine Translation of Rare Words with Subword Units". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. DOI: 10.18653/v1/P16-1162. URL: https://aclanthology.org/P16-1162/.

[30] Zellig Harris. "Distributional structure". In: *Word* 10.2-3 (1954), pp. 146–162. DOI: 10.1007/978-94-009-8467-7_1. URL: https://link.springer.com/chapter/10.1007/978-94-009-8467-7_1.

[31] Raia Hadsell, Sumit Chopra, and Yann LeCun. "Dimensionality Reduction by Learning an Invariant Mapping". In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17–22 June 2006, New York, NY, USA*. IEEE Computer Society, 2006, pp. 1735–1742. DOI: 10.1109/CVPR.2006.100. URL: https://doi.org/10.1109/CVPR.2006.100.

[32] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. "Representation Learning with Contrastive Predictive Coding". In: *CoRR* abs/1807.03748 (2018). arXiv: 1807.03748. URL: http://arxiv.org/abs/1807.03748.

[33] Nils Reimers and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. DOI: 10.18653/v1/D19-1410. URL: https://aclanthology.org/D19-1410/.

[34] Long Ouyang et al. “Training Language Models to Follow Instructions with Human Feedback”. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS ’22. Red Hook, NY, USA: Curran Associates Inc., 2022.

[35] Jason Wei et al. “Finetuned Language Models Are Zero-Shot Learners”. In: *International Conference on Learning Representations* (2022).

[36] Shangyu Wu et al. “Retrieval-Augmented Generation for Natural Language Processing: A Survey”. In: *Artificial Intelligence Review* (2026). DOI: 10.1007/s10462-026-11605-7.

[37] Shunyu Yao et al. “ReAct: Synergizing Reasoning and Acting in Language Models”. In: *ICLR*. OpenReview.net, 2023. URL: https://openreview.net/forum?id=WE_vluYUL-X.

[38] Hugging Face. *Agents Course — Unit 1: Introduction to Agents*. https://huggingface.co/learn/agents-course/unit1. 2024.

[39] Tomas Mikolov et al. “Distributed Representations of Words and Phrases and their Compositionality”. In: *Advances in Neural Information Processing Systems*. Vol. 26. Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

[40] Jeffrey Pennington, Richard Socher, and Christopher Manning. “GloVe: Global Vectors for Word Representation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162/.

[41] Liang Wang et al. *Text Embeddings by Weakly-Supervised Contrastive Pre-training*. 2024. arXiv: 2212.03533 [cs.CL]. URL: https://arxiv.org/abs/2212.03533.

[42] Zehan Li et al. *Towards General Text Embeddings with Multi-stage Contrastive Learning*. Aug. 2023. DOI: 10.48550/arXiv.2308.03281.

[43] Colin Raffel et al. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”. In: *Journal of Machine Learning Research* 21.140 (2020).

[44] Jianlin Su et al. “RoFormer: Enhanced Transformer with Rotary Position Embedding”. In: *Neurocomputing* 568 (2024), p. 127063. DOI: 10.1016/j.neucom.2023.127063.

[45] Jianlv Chen et al. "M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation". In: *Findings of the Association for Computational Linguistics: ACL 2024*. Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 2318–2335. DOI: 10.18653/v1/2024.findings-acl.137.

[46] Liang Wang et al. "Multilingual E5 Text Embeddings: A Technical Report". In: *arXiv preprint arXiv:2402.05672* (2024).

[47] Zach Nussbaum et al. *Nomic Embed: Training a Reproducible Long Context Text Embedder*. 2024. arXiv: 2402.01613 [cs.CL].

[48] Gabriel de Souza P. Moreira et al. "Improving Text Embedding Models with Positive-aware Hard-negative Mining". In: CIKM '25. New York, NY, USA: Association for Computing Machinery, 2025, pp. 2169–2178. DOI: 10.1145/3746252.3761254. URL: https://doi.org/10.1145/3746252.3761254.

[49] Google DeepMind. *Generalizable Embeddings from Gemini*. 2025. arXiv: 2503.07891 [cs.CL].

[50] OpenAI. *New Embedding Models and API Updates*. https://openai.com/blog/new-embedding-models-and-api-updates. 2024.

[51] Microsoft. *What Is Azure OpenAI Service?* https://azure.microsoft.com/it-it/pricing/details/azure-openai/. Accessed: 2025-04-30. 2024.

[52] Giorgia Bolognesi et al. *LARAG: Link-Aware Retrieval Strategy for RAG Systems in Hyperlinked Technical Documentation*. Tech. rep. Preprint. Rulex Innovation Labs s.r.l. / IMATI-CNR / University of Genova (DIMA, DIBRIS), Feb. 2026.

[53] Ziwei Ji et al. "Survey of Hallucination in Natural Language Generation". In: *ACM Computing Surveys* 55.12 (2023).

[54] Jeff Johnson, Matthijs Douze, and Hervé Jégou. "Billion-Scale Similarity Search with GPUs". In: *IEEE Transactions on Big Data* 7.3 (2019), pp. 535–547.

[55] Jason Wei et al. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". In: *Advances in Neural Information Processing Systems*. Vol. 35. Curran Associates, Inc., 2022, pp. 24824–24837. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

[56] Rulex. *Rulex Platform Documentation, Version 1.4.x*. https://doc.rulex.ai/docs/v14/index.html. Accessed: 2025-04-30. 2025.

[57] Roger Mitton. *Spelling in the Birkbeck Corpus*. Tech. rep. Available from the Oxford Text Archive. Birkbeck College, University of London, 1985. URL: https://titan.dcs.bbk.ac.uk/~roger/missp.dat.

[58] Wikipedia contributors. *Wikipedia: Lists of Common Misspellings*. https://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings. Accessed: 2024. Compiled list of recurring typographical errors observed in Wikipedia edits, maintained collaboratively by the Wikipedia community. 2024. URL: https://titan.dcs.bbk.ac.uk/~roger/wikipedia.dat.

[59] Robert Edward Lewand. *Cryptological Mathematics*. Letter frequency table reproduced at https://mathcenter.oxford.emory.edu/site/math125/englishLetterFreqs/. Mathematical Association of America, 2000.

[60] Tri Nguyen et al. "MS MARCO: A Human Generated MAchine Reading COmprehension Dataset". In: *CoCo@NIPS*. Vol. 1773. CEUR Workshop Proceedings. CEUR-WS.org, 2016. URL: https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf.

[61] Tom Kwiatkowski et al. "Natural Questions: A Benchmark for Question Answering Research". In: *Transactions of the Association for Computational Linguistics* 7 (2019), pp. 452–466. DOI: 10.1162/tacl_a_00276. URL: https://aclanthology.org/Q19-1026/.

[62] Harrison Chase. *LangChain*. https://github.com/langchain-ai/langchain. Accessed: 2025-04-30. 2024.

[63] Chroma Core. *Chroma: The Open-Source Embedding Database*. https://www.trychroma.com. Accessed: 2025-04-30. 2023.

[64] Tianyi Zhang et al. "BERTScore: Evaluating Text Generation with BERT". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020*. OpenReview.net, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

Image Sources

[65] Wikimedia Commons. *Illustration of Unit Circles (Actually Unit "Balls") for Various p-Norms*. File: `Vector_norms.svg`. Licensed under CC BY-SA 3.0. Accessed: 2026. 2006. URL: https://en.wikipedia.org/wiki/Lp_space.

[66] Stefan Sommer, Tom Fletcher, and Xavier Pennec. "Introduction to Differential and Riemannian Geometry". In: *Riemannian Geometric Statistics in Medical Image Analysis*. Elsevier, 2020. Chap. 1, pp. 3–37. DOI: 10.1016/b978-0-12-814725-2.00008-x.

[67] Emma Aspland et al. "Modified Needleman-Wunsch algorithm for clinical pathway clustering". In: *Journal of biomedical informatics* 115 (Dec. 2020), p. 103668. DOI: 10.1016/j.jbi.2020.103668.

[68] Lilian Weng. *From Autoencoder to Beta-VAE*. Lil'Log. Accessed: 2026. Aug. 2018. URL: https://lilianweng.github.io/posts/2018-08-12-vae/.

[69] Keenan Crane. *Illustration of Latent Manifold via Autoencoder*. Mathstodon. Accessed: 2026. 2024. URL: https://mathstodon.xyz/@keenancrane/media.

[70] Riccardo Stagliandò. *L'ia e altre cose spiegate benissimo*. Substack, *Artificiale*. Accessed: 2026. 2023. URL: https://riccardostagliano.substack.com/p/122-lia-e-altre-cose-spiegate-benissimo.

[71] Adam Bohr. *A Complete Guide to BERT with Code*. https://towardsdatascience.com/a-complete-guide-to-bert-with-code-9f87602e4a11/. Towards Data Science. 2024.