BÀI TẬP 2

TÌM KIẾM VÀ ĐÁNH GIÁ THÔNG TIN HỌC THUẬT

Ứng dụng của Large Language Models (LLMs) trong phân tích dữ liệu bảng

I. Mục tiêu

Báo cáo này ghi lại quá trình tìm kiếm, chọn lọc và đánh giá thông tin học thuật từ nhiều loại nguồn khác nhau. Mục tiêu của bài là rèn luyện khả năng xác định chủ đề nghiên cứu, tìm tài liệu từ các cơ sở dữ liệu học thuật, tạp chí khoa học, sách chuyên khảo và nguồn mở Internet; đồng thời đánh giá độ tin cậy của từng nguồn dựa trên tác giả, nơi công bố, phương pháp nghiên cứu, mức độ cập nhật và khả năng kiểm chứng.

II. Lựa chọn chủ đề và phạm vi tìm kiếm

Chủ đề: Ứng dụng của Large Language Models (LLMs) trong phân tích dữ liệu bảng (Tabular Data).

Lý do chọn chủ đề: Dữ liệu bảng là dạng dữ liệu phổ biến trong doanh nghiệp, tài chính, y tế, giáo dục và nhiều hệ thống quản trị. Trong khi các mô hình cây như Random Forest, XGBoost hoặc LightGBM vẫn rất mạnh trên dữ liệu bảng, sự phát triển của LLMs và foundation models đặt ra câu hỏi liệu các mô hình ngôn ngữ lớn có thể hỗ trợ, bổ sung hoặc thay thế một phần quy trình phân tích dữ liệu bảng hay không.

Phạm vi tìm kiếm: Báo cáo tập trung vào các tài liệu công bố trong giai đoạn 2022-2026, ưu tiên các nghiên cứu liên quan đến LLMs, foundation models, mô hình học máy truyền thống và ứng dụng thực tế trong Data Science.

III. Quá trình tìm kiếm thông tin

- Cơ sở dữ liệu học thuật: Google Scholar, arXiv, Semantic Scholar. Từ khóa sử dụng gồm: "LLM for Tabular Data", "Tabular Foundation Models", "Language Modeling on Tabular Data", "Tabular Prediction".
- Tạp chí và hội nghị khoa học: ICLR, NeurIPS, IEEE, JMLR và các kho bài báo học thuật có phản biện hoặc tiền ấn phẩm được cộng đồng nghiên cứu sử dụng rộng rãi.
- Sách chuyên khảo: Sách về Machine Learning và Data Science từ các nhà xuất bản uy tín như O'Reilly hoặc Springer để đối chiếu kiến thức nền.
- Nguồn mở Internet: Kaggle, báo cáo công nghiệp, blog kỹ thuật của các tổ chức AI và bài viết cộng đồng. Các nguồn này được dùng như nguồn bổ trợ, không thay thế cho tài liệu học thuật.

IV. Bảng đánh giá độ tin cậy của nguồn

Các nguồn thông tin dưới đây được đánh giá theo các tiêu chí: uy tín tác giả, nơi công bố, phương pháp nghiên cứu, mức độ cập nhật, khả năng kiểm chứng và mức độ phù hợp với chủ đề.

STT	Nguồn/Tài liệu	Loại nguồn	Đánh giá độ tin cậy	Xếp hạng

01	Fang et al. (2024) - Large Language Models on Tabular Data: Prediction, Generation, and Understanding	Bài báo khảo sát / arXiv	Khảo sát trực tiếp về LLMs với dữ liệu bảng; bao quát dự đoán, sinh dữ liệu, hỏi đáp và hiểu bảng.	Rất cao
02	Ruan et al. (2024) - Language Modeling on Tabular Data	Bài báo khảo sát / arXiv	Tổng hợp nền tảng và kỹ thuật language modeling cho dữ liệu bảng; giúp hiểu bức tranh nghiên cứu chung.	Rất cao
03	Gardner, Perdomo and Schmidt (2024) - Large Scale Transfer Learning for Tabular Data via Language Modeling	Bài báo nghiên cứu / arXiv	Giới thiệu TabuLa-8B cho dự đoán dữ liệu bảng; có quy trình huấn luyện và đánh giá trên nhiều bộ dữ liệu.	Cao
04	Hollmann et al. (2022) - TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second	Bài báo nghiên cứu / ICLR/arXiv	Foundation model quan trọng cho dữ liệu bảng; có mã nguồn và thực nghiệm trên nhiều bộ dữ liệu nhỏ.	Rất cao
05	Grinsztajn, Oyallon and Varoquaux (2022) - Why do tree-based models still outperform deep learning on typical tabular data?	Bài báo hội nghị / NeurIPS	Thực nghiệm quy mô lớn; là nguồn đối chứng khi so sánh deep learning/LLMs với mô hình cây.	Rất cao
06	Borisov et al. (2022) - Deep Learning on Tabular Data: A Survey	Bài báo khảo sát	Tổng quan về deep learning trên dữ liệu bảng; hữu ích để đặt LLMs vào bối cảnh phương pháp trước đó.	Cao
07	Géron (2022) - Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow	Sách chuyên khảo	Sách nền tảng uy tín, phù hợp để đối chiếu kiến thức về pipeline và đánh giá mô hình.	Rất cao
08	Kaggle - State of Data Science and Machine Learning Survey	Báo cáo cộng đồng / nguồn mở	Phản ánh xu hướng công cụ trong cộng đồng Data Science; thực tế nhưng cần đối chiếu với nguồn học thuật.	Khá cao
09	McKinsey Global Institute - The Economic Potential of Generative AI	Báo cáo công nghiệp	Giúp hiểu tác động kinh tế và ứng dụng doanh nghiệp của GenAI; nên dùng như nguồn bối cảnh.	Khá cao
10	Google DeepMind Blog / Towards Data Science technical posts	Blog kỹ thuật / bài viết Internet	Cập nhật nhanh, dễ hiểu và có tính ứng dụng; mức độ phản biện thấp hơn, dùng làm nguồn phụ.	Trung bình

V. Nhận xét tổng quan

Qua bảng đánh giá, có thể thấy các bài báo khảo sát và nghiên cứu học thuật có độ tin cậy cao nhất vì trình bày rõ phương pháp, phạm vi nghiên cứu, giới hạn của mô hình và có khả năng kiểm chứng. Các bài nghiên cứu như TabPFN hoặc TabuLa-8B giúp hiểu rõ cách foundation models được áp dụng vào dữ liệu bảng. Trong khi đó, sách chuyên khảo giúp củng cố kiến thức nền về machine learning và quy trình đánh giá mô hình.

Các nguồn như Kaggle, McKinsey, blog kỹ thuật và bài viết cộng đồng có lợi thế về tính cập nhật và góc nhìn thực tế. Tuy nhiên, chúng thường không có quy trình phản biện học thuật nghiêm ngặt như bài báo khoa học, nên chỉ nên dùng làm nguồn phụ hoặc để bổ sung bối cảnh ứng dụng.

VI. Danh mục tài liệu tham khảo theo định dạng Harvard

1. Fang, X., Xu, W., Tan, F.A., Zhang, J., Hu, Z., Qi, Y., Nickleach, S., Socolinsky, D., Sengamedu, S. and Faloutsos, C. (2024) Large Language Models (LLMs) on Tabular Data: Prediction, Generation, and Understanding - A Survey. arXiv:2402.17944.
2. Ruan, Y., Lan, X., Ma, J., Dong, Y., He, K. and Feng, M. (2024) Language Modeling on Tabular Data: A Survey of Foundations, Techniques and Evolution. arXiv:2408.10548.
3. Gardner, J., Perdomo, J.C. and Schmidt, L. (2024) Large Scale Transfer Learning for Tabular Data via Language Modeling. arXiv:2406.12031.
4. Hollmann, N., Müller, S., Eggensperger, K. and Hutter, F. (2022) TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. arXiv:2207.01848.
5. Grinsztajn, L., Oyallon, E. and Varoquaux, G. (2022) Why do tree-based models still outperform deep learning on typical tabular data? Advances in Neural Information Processing Systems.
6. Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M. and Kasneci, G. (2022) Deep learning on tabular data: A survey. IEEE Transactions on Neural Networks and Learning Systems.
7. Géron, A. (2022) Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. 3rd edn. Sebastopol: O'Reilly Media.

8. Kaggle (n.d.) State of Data Science and Machine Learning Survey. Available at: https://www.kaggle.com/ (Accessed: 7 June 2026).
9. McKinsey Global Institute (2023) The Economic Potential of Generative AI: The Next Productivity Frontier. Available at: https://www.mckinsey.com/ (Accessed: 7 June 2026).
10. Google DeepMind (n.d.) Technical blog and AI research updates. Available at: https://deepmind.google/ (Accessed: 7 June 2026).