Balancing Explainability and Performance:

A Cross-Domain Comparison of Pre-Trained AI Models

Robert McCoy

Indiana Wesleyan University

Model Development (AIML-501-01A)

Professor L Lounge

July 14, 2025

Balancing Explainability and Performance:

A Cross-Domain Comparison of Pre-Trained AI Models

Introduction

As artificial intelligence (AI) applications proliferate across industries, the need for explainability has moved from an academic concern to an operational imperative. Explainability refers to the ability to understand, interpret, and trust the decisions made by AI models—especially in domains where human lives, finances, or legal judgments may be impacted (Barredo Arrieta et al., 2020). Model opacity, often called the "black box" problem, impedes stakeholder trust, limits accountability, and raises ethical concerns.

Pre-trained models, which are widely adopted for their state-of-the-art performance, present unique trade-offs between accuracy, speed, model size, and transparency. Some models boast tremendous predictive capability but offer little interpretability; others prioritize traceability and simplicity but sacrifice complexity. Navigating these trade-offs is central to building responsible AI systems.

This paper compares representative models from three domains—NLP/Generative AI, Computer Vision, and Tabular Data—highlighting their strengths, weaknesses, and explainability profiles. In addition, recent advances in data-efficient fine-tuning techniques, particularly those developed by Microsoft (Sun et al., 2025), are examined for their potential contributions to transparency, version control, and auditability.

Methodology

To ensure domain diversity, this analysis focuses on four models:

- GPT-3.5: A transformer-based generative language model known for its impressive fluency and general-purpose capabilities (Brown et al., 2020).
- BERT-base: A bidirectional transformer optimized for language understanding and classification (Devlin et al., 2019).
- EfficientNet-B0: A convolutional neural network (CNN) model scaled for efficient image classification (Tan & Le, 2019).
- XGBoost: A scalable, tree-based ensemble algorithm widely used in structured data applications (Chen & Guestrin, 2016).

Model characteristics—including number of parameters, benchmark accuracy, inference latency, and explainability—were sourced from peer-reviewed publications, model repositories, and recent technical papers. Inferences about transparency and interpretability were also informed by explainability tools and known industry practices (Molnar, 2022).

Decision Matrix

Table 1

Comparison of Pre-Trained Models Across Domains

Model	Domain	Size	Accuracy	Inference Speed	Explainability	Use Case Fit
GPT-3.5	NLP / Generative AI	175B+ params	Slow	✕ Slow	✕ Low (Opaque)	Text generation, creative writing
BERT-base	NLP / Understanding	110M params	~82% (GLUE Benchmark)	✅ Fast	⚠ Moderate	Sentiment analysis, text classification
EfficientNet-B0	Computer Vision	~5M params	~77% (ImageNet Top-1)	✅ Fast	⚠ Moderate	Mobile vision tasks, object detection
XGBoost	Tabular Data	Low (KB–MB)	Very High (Kaggle Proven)	✅✅ Very Fast	✅ High (Transparent)	Structured risk modeling, business analytics

Note: Matrix developed in correspondence with AI assistant (McCoy, 2025).

Analysis and Recommendations

Each model addresses different challenges, and their trade-offs should be interpreted in context.

- GPT-3.5, while exceptional in generative applications, is costly and inherently opaque. Its layered transformer architecture and emergent behaviors are difficult to interpret (Bommasani et al., 2021), making it risky for use in regulated industries.
- BERT-base offers a more compact and faster alternative, with moderate explainability when paired with probing methods like SHAP or attention visualization. It is suitable for applications requiring both speed and decent transparency, such as text classification or customer feedback analysis.

- EfficientNet-B0 is highly optimized for speed and parameter efficiency in computer vision but suffers from low native explainability—CNNs are notoriously hard to interpret without tools like Grad-CAM (Selvaraju et al., 2017).
- XGBoost, by contrast, remains highly interpretable and scalable for structured data. It integrates well with explainability tools such as feature importance plots, decision tree visualizations, and SHAP values (Molnar, 2022). This makes it a preferred model for domains requiring audit trails and human oversight.

Microsoft's Reinforced Fine-Tuning (Optional Efficiency Amplifier)

Microsoft's recent work on reinforced pre-training (Sun et al., 2025) introduces two innovations: adaptive difficulty targeting and rollout replay, both of which increase training data efficiency and reduce fine-tuning time by up to 65%. While this research primarily targets training cost reduction, it also improves model auditability by making training iterations more manageable and traceable. These changes could be valuable in environments where model explainability depends on transparency throughout the development pipeline—not just in inference outputs.

Conclusion

This cross-domain evaluation highlights that no single model optimally balances accuracy, speed, and explainability across all tasks. GPT-3.5 delivers state-of-the-art generative capacity but lacks transparency. BERT and EfficientNet offer balanced performance but moderate interpretability. XGBoost excels in both speed and explainability for tabular contexts.

Future development in AI, including data-efficient training strategies such as Microsoft's reinforced fine-tuning, may shift the trade-offs by lowering the barriers to retraining, debugging, and version control. However, responsible AI implementation demands more than high performance—it requires deliberate model selection that prioritizes explainability and ethical alignment in the chosen domain.

References

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *https://arxiv.org/abs/2108.07258*

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. ht*tps://arxiv.org/abs/1810.04805*

McCoy, R. (2025). *Decision matrix from professional correspondence with AI advisor*. Personal communication, June 28, 2025.

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). https://christophm.github.io/interpretable-ml-book/

Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., & Batra, D. (2016). *Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization*. https://arxiv.org/abs/1610.02391v1

Sun, Y., Hu, J., Xu, H., Qiu, L., & Lin, Z. (2025 Improving Data Efficiency for LLM Reinforcement Fine-tuning Through Difficulty-targeted Online Data Selection and Rollout Replay. *https://arxiv.org/abs/2506.05316*

Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *https://arxiv.org/abs/1905.11946*