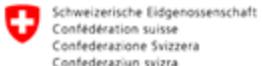


# DIGITAL FINANCE

This project has received funding from the  
Horizon Europe research and innovation  
programme under the Marie Skłodowska-  
Curie Grant Agreement No. 101119635



Schweizerische Eidgenossenschaft  
Confédération suisse  
Confederazione Svizzera  
Confederaziun svizra

State Secretariat for Education,  
Research and Innovation SERI



Funded by  
the European Union



# Reading

---

Bilal, A., Ebert, D., & Lin, B. (2025). *LLMs for explainable AI: A comprehensive survey*. arXiv.  
<https://arxiv.org/abs/2504.00125>



# LLMs for XAI

Bilal, A., Ebert, D., & Lin, B. (2025). LLMs for explainable AI: A comprehensive survey. arXiv.  
<https://arxiv.org/abs/2504.00125>



DIGITAL

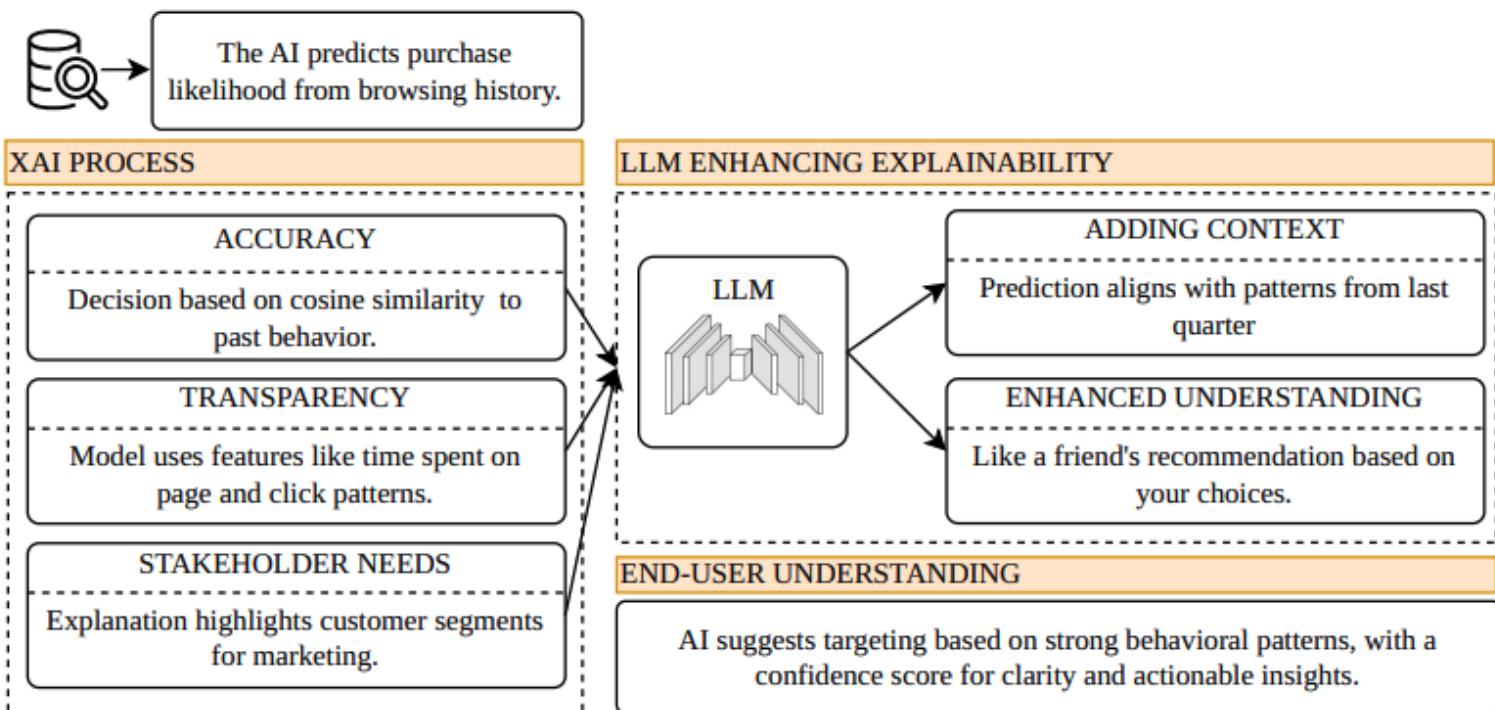


Funded by  
the European Union

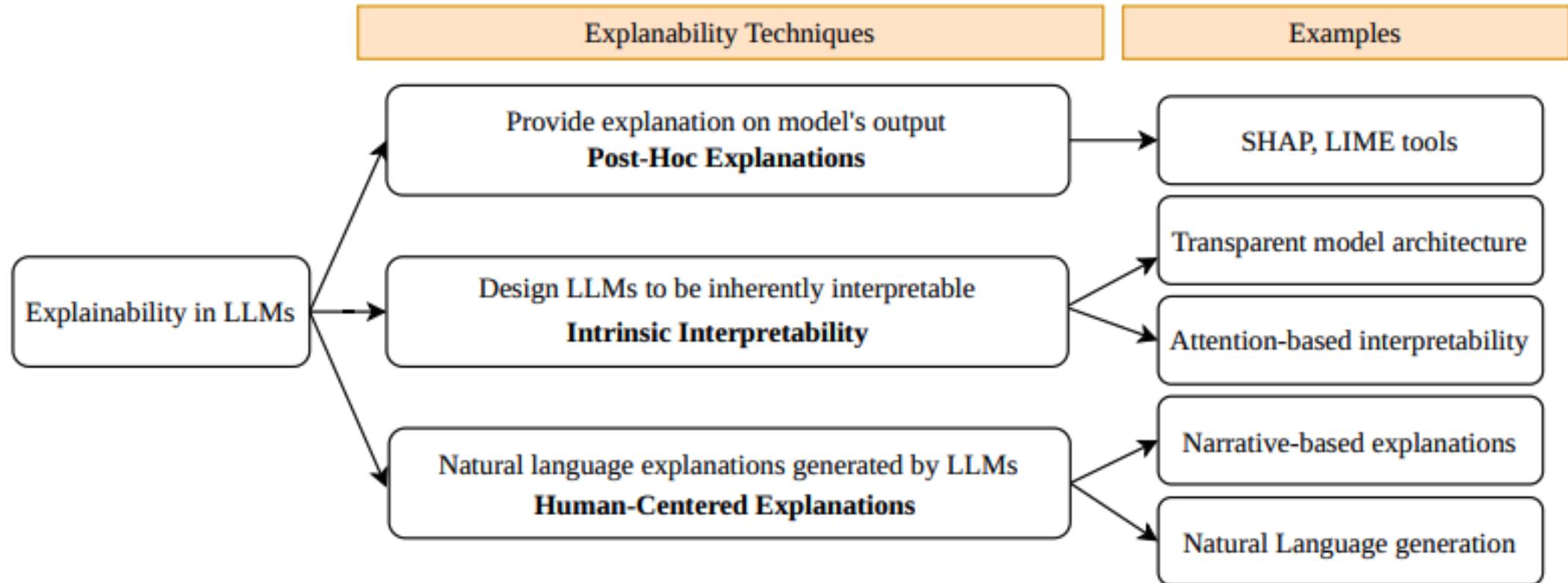
# A Comprehensive Survey

---

- A **unified conceptual framework** for how LLMs are used in Explainable AI, plus a structured overview of **evaluation, datasets, applications, and open challenges**.



DIGITAL



Bilal, A., Ebert, D., & Lin, B. (2025). *LLMs for explainable AI: A comprehensive survey*. arXiv. <https://arxiv.org/abs/2504.00125>



# How do we evaluate LLM-generated explanations?

---

Category	Metric	Description	Example
Qualitative	Comprehensibility and human understanding [86, 126]	Ease of understanding and clarity in delivering the model's reasoning to humans	Explaining sentiment analysis by highlighting key phrases, such as "masterful direction" and "innovative storytelling"
	Controllability [31, 36]	Interactivity and adjustability of explanations, enabling users to provide feedback to refine explanations	Users highlighting unclear parts of the explanation for improvement
Quantitative	Faithfulness [23]	Accuracy in representing the model's internal decision-making process, including causal relationships and alignment with training data	A medical diagnosis system showing specific patient features and their weights that informed the decision
	Plausibility [64, 105]	Logical coherence and domain consistency, ensuring alignment with established knowledge	A climate prediction model providing explanations consistent with meteorological principles



# Benchmark datasets for explanation learning



DIGITAL

Dataset	Focus/Use	Example
<b>e-SNLI</b> [28]	Human-written explanations for tasks to find logical relation.	Sentence and label pairs, such as "entailment," "contradiction," or "neutral," e.g., "A dog is a type of animal, and running through a park outside."
<b>CoS-E</b> [116]	Multichoice common-sense reasoning with explanations.	Example: "Why wear sunglasses?", pair with the answer: "To block the sun," with explanation: "Sunglasses protect eyes from bright sunlight."
<b>ECQA</b> [5]	Commonsense reasoning with detailed explanations.	"Why lock a car?" with answer: "To prevent theft," with explanation: "Locks secure doors, preventing unauthorized access."
<b>WorldTree</b> [66]	Scientific reasoning using structured explanation graphs.	Example: "Why does a metal spoon feel hotter than plastic in hot water?" with the answer: "Metal conducts heat efficiently," and explanation: Graph with domain facts, such as "Metal is a conductor."
<b>OpenBookQA</b> [99]	Elementary science questions with knowledge-based explanations.	Example: "Which lets heat travel best?" with answer: "Steel spoon," and explanation: "Metal is a thermal conductor, and steel is made of metal."
<b>XplainLLM</b> [35]	Factual reasoning with knowledge graphs in triplet form.	Example: "Why did people dance to music?" with answer: "Enjoyable," and explanation: Knowledge graph linking "enjoyable" with relevant reasoning.
<b>RAGBench</b> [53]	Retrieval-Augmented Generation with explainability for specific domains.	Example: "How to reset a device?" with explanation: Retrieved evidence from a user manual supports the generated response.
<b>HateXplain</b> [96]	Hate speech detection with unbiased explanations.	Example: A flagged post with highlighted words or phrases explaining the "hate speech" classification.

# Challenges and Limitations

---

## Sensitive data & privacy

- LLM explanations often verbalize sensitive features (medical, financial, behavioral).
- Natural language explanations increase the risk of information leakage compared to numeric attributions.
- Tension between right to explanation and data minimization / confidentiality.
- Over-explaining can violate legal and ethical privacy constraints.



DIGITAL

# Challenges and Limitations

---

## Cultural and societal norms

- Explanations are **not culturally neutral**; meaning depends on social context.
- LLMs reflect **dominant cultural and linguistic norms** from training data.
- A single explanation may be acceptable to one group but **misleading or offensive** to another.
- Risk of **misaligned trust** in cross-cultural or global deployments.



DIGITAL

# Challenges and Limitations

---

## Multi-source data conflicts

- AI decisions often combine **heterogeneous data sources** with unequal reliability.
- LLMs tend to **smooth over conflicts** and present a single coherent narrative.
- Narrative coherence may **mask weak, biased, or irrelevant signals**.
- Explanations can falsely imply **causal legitimacy** of all data sources.



DIGITAL

# Challenges and Limitations

---

## Model complexity

- Deep models rely on **non-linear, distributed representations** that are hard to trace.
- LLMs generate **descriptions of reasoning**, not the model's actual computations.
- Creates an **illusion of transparency** without true interpretability.
- Risk of over-trust in explanations that are only approximations.



DIGITAL

# Challenges and Limitations

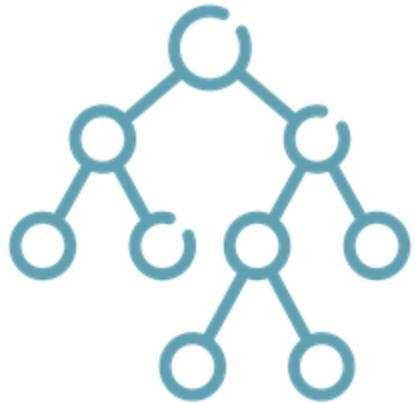
---

## Bias and fairness in LLMs

- LLMs inherit **social, linguistic, and representation biases** from training data.
- Bias affects **both decisions and explanations**.
- Explanations may **legitimize or normalize unfair outcomes**.
- Biased explanations are especially dangerous because they sound **plausible and justified**.



DIGITAL



# DIGITAL



**Funded by  
the European Union**



DIGITAL

*This project has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101119635*

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or Horizon Europe: Marie Skłodowska-Curie Actions. Neither the European Union nor the granting authority can be held responsible for them.