

# Enhancing model transparency: Effects of DK Integration for Conversational XAI

Isabel Feustel | Ulm University



Supported by  
**Financial support programmes  
for early career researchers**  
Graduate & Professional Training Center  
Ulm University

# About me



## Education

M. Sc. Media Informatics

PhD Student at Ulm  
University



## Research Interests

Dialogue Systems

Explainable AI

Human Computer  
Interaction



## Side facts

I love music

I'm addicted to video games

I owned pet rats

# Outline



Motivation



Foundations of XAI



Concept for Domain Knowledge Integration



Automatic Generation of Structured Knowledge

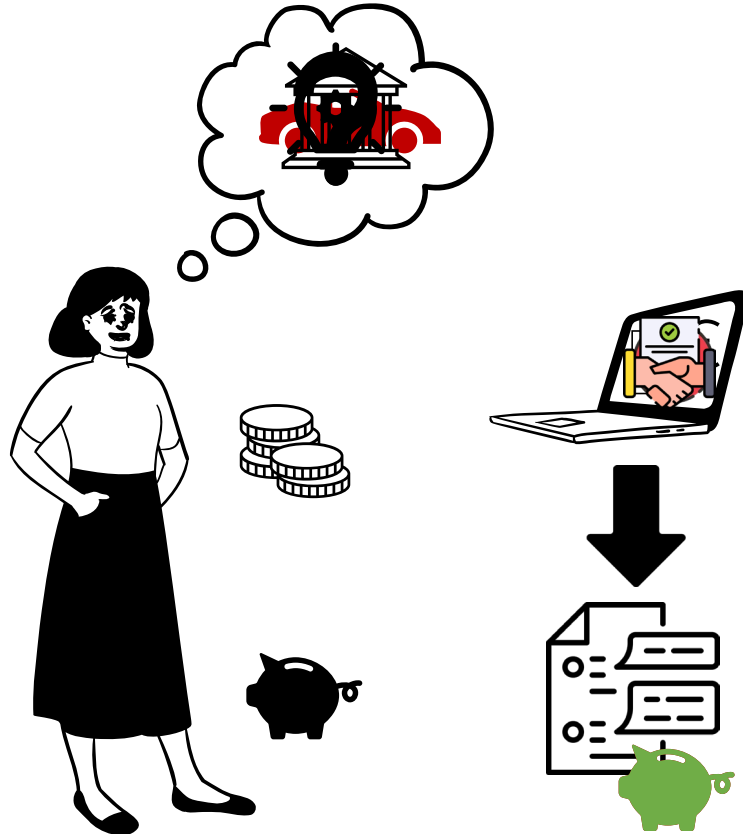


Evaluation



Conclusion

# A Conversation with AI: Understanding Model Decisions



Regulations for transparent AI  
EU AI Intelligence Act



Help users understanding  
AI behavior



Empowerment through  
understanding



Gain trust in AI systems



# Conversation as a basis



Natural way of explanations



Information can be splitted (No overload)



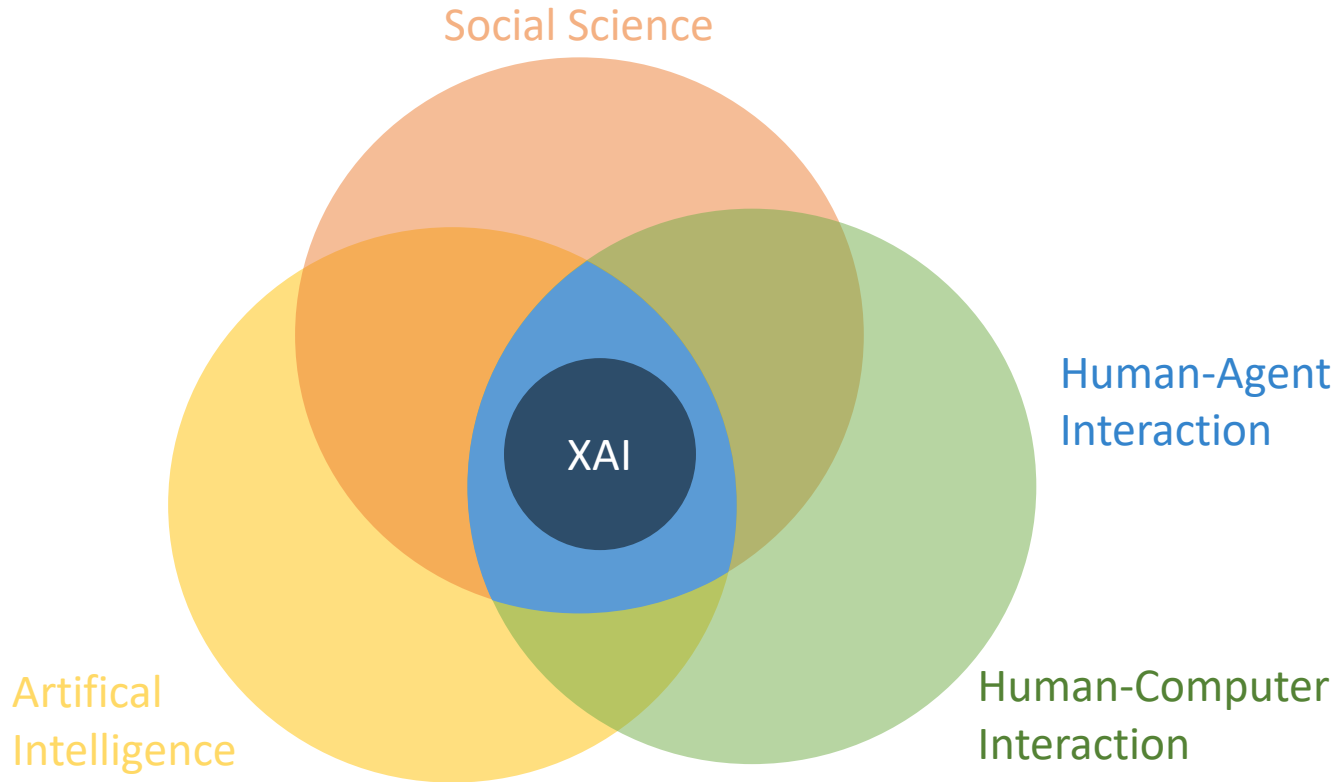
User can ask questions for clarification



Personalized experience



# Transparency is the key – How do we get there?



# XAI Foundations: Model Explainability



Post-Hoc Methods



Interpretable models  
(white boxes)



# XAI Foundations: Explanation Types



## Local Explanation

Why specific prediction?  
Instance level



## Global Explanation

How does model behave overall?  
Model level



# XAI Foundations: Explanation Types



## Local Explanation

Why specific prediction?  
Instance level



## Feature Importance

What feature influenced the outcome?



## Counterfactual Explanation

What changes the outcome?



# Limitations of XAI



Non conversational explanations



Main focus on expert users



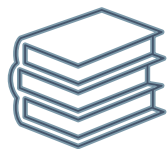
Limited to data and model only



# Limitations of XAI



Limited to data and model only



Integrate knowledge



# Outline



Motivation



Foundations of XAI



Concept for Domain Knowledge Integration



Automatic Generation of Structured Knowledge



Evaluation



Conclusion

# Example Scenario Credit Application

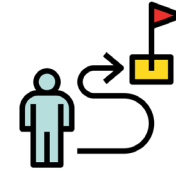
- Prediction: Would you be approved for a credit loan?



savings



checking  
account



purpose



duration



amount

# Unified Explanations

## XAI

For the given case, purpose is the most influential feature.

Model

Data

Credit applications for used cars have a higher acceptance rate (83%) compared to those for new cars (62%).

In this case, the purpose influenced the loan approval decision.

This can be explained by the disparity between used car and new car loans, with acceptance rates of 83% and 62% respectively.

Further, used cars, primarily viewed as a means of transportation, often involve lower loan amounts due to their lower purchase price.

In contrast, new cars, frequently seen as status symbols, may be perceived as higher-risk purchases, potentially influenced by factors beyond essential transportation needs.

## External Knowledge

Used cars typically cost less than new cars. This leads to smaller loan amounts, reducing the lender's risk.

Domain Knowledge

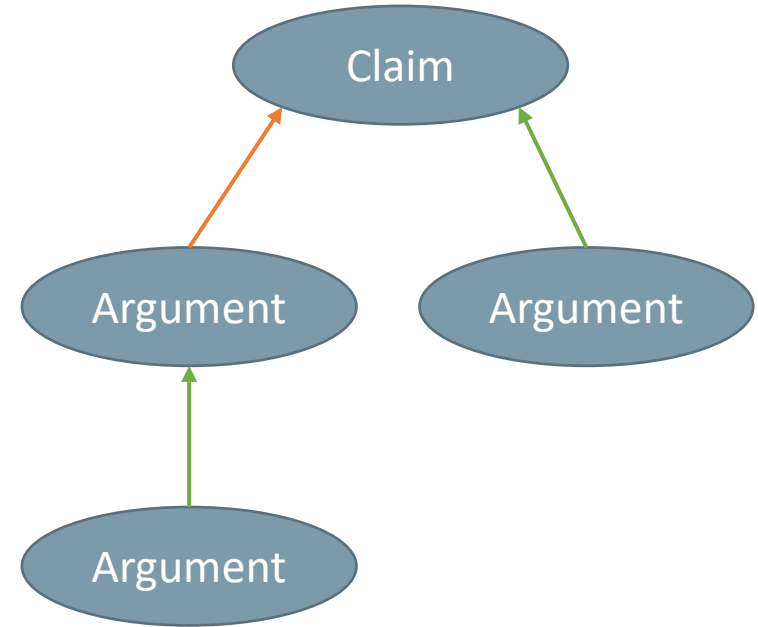
World Knowledge

New cars often serve as status symbols, while used cars are primarily viewed as a means of transportation.

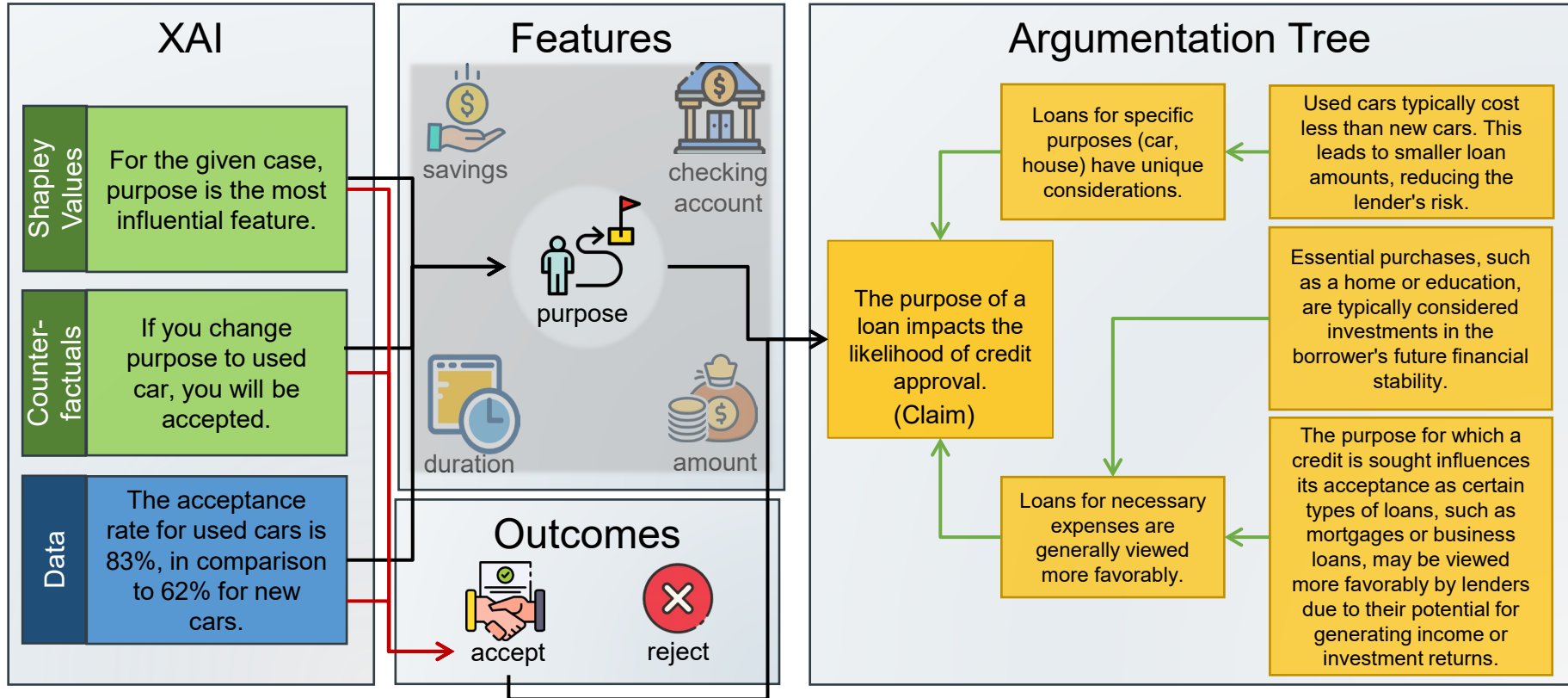


# Leveraging Argumentation to Enhance XAI Explanations

- Explanations and reasoning of humans are argumentative (*Mercier et al., 2011; Antaki et al., 1992*)
- Argumentation frameworks as basis (*Stab et. al, 2014*)
  - Argument Components
  - Argumentative Relations
- Argumentation Trees offer dialogical access



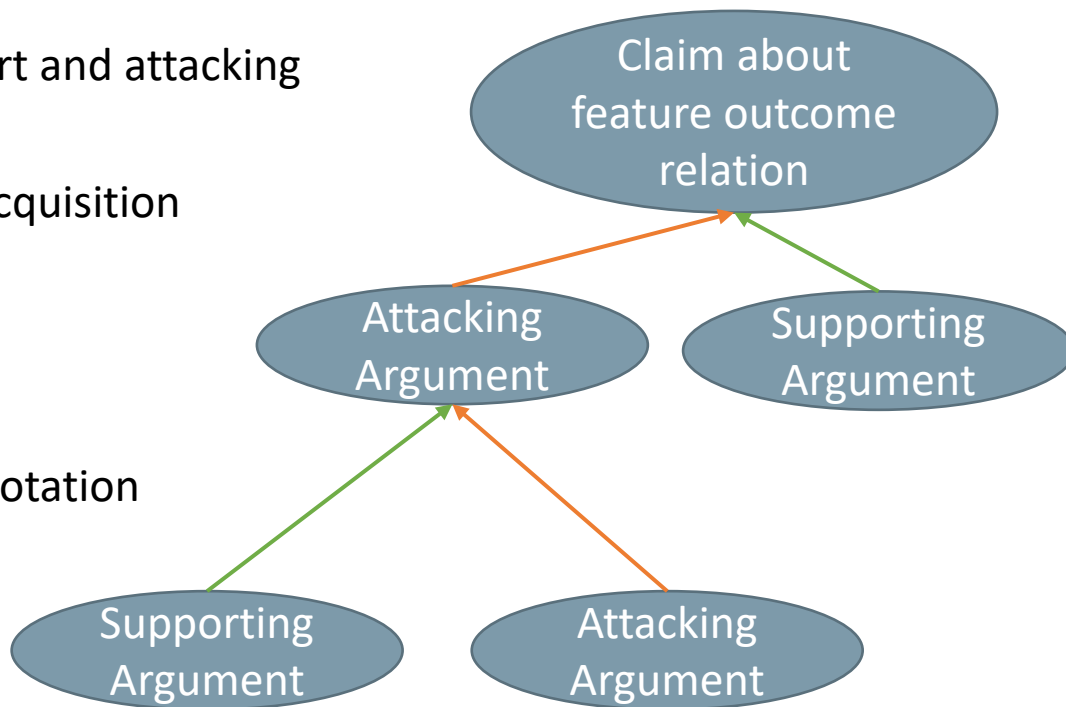
# The Bridge – Linking XAI and Argumentation





# Building the Knowledge Base: Constructing Argument Trees

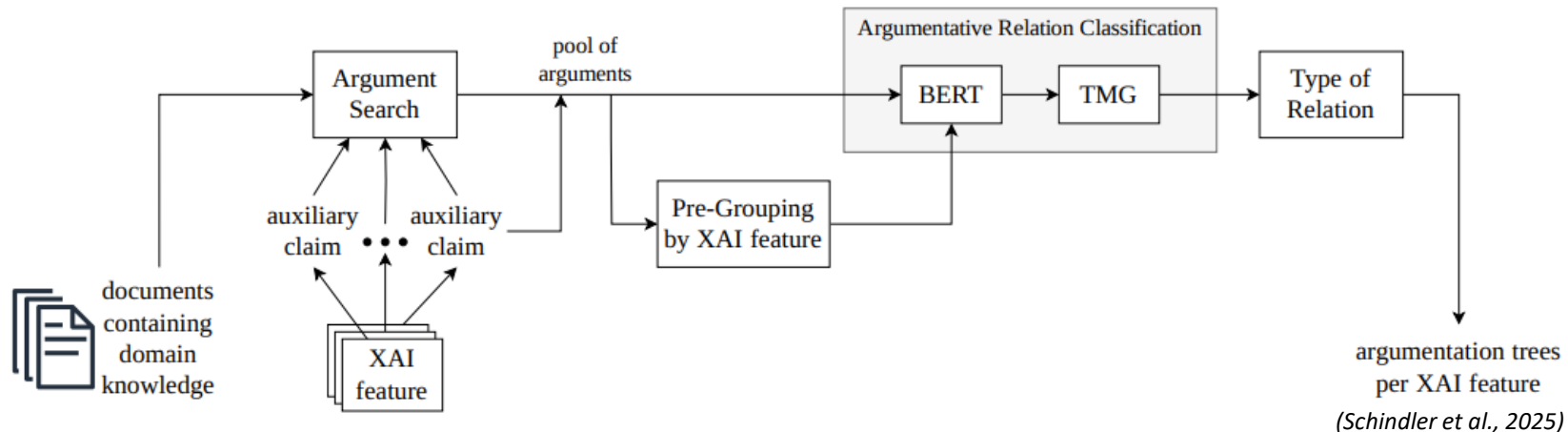
- Argumentation Tree with support and attacking arguments
- 3 Types of domain knowledge acquisition
  - Handcrafted
  - LLM generated
  - Pipeline for generation
- All 3 types included manual annotation processes



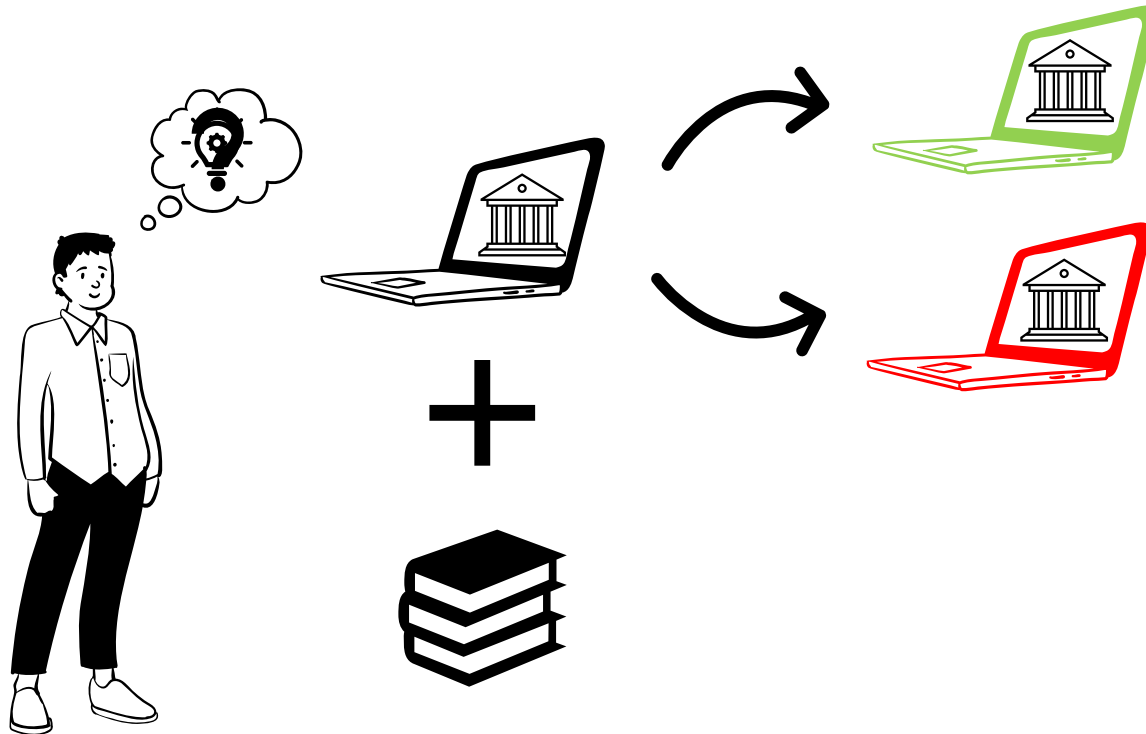
(Schindler et al., 2025)

# Automatic generation of argumentation structures for conversational XAI

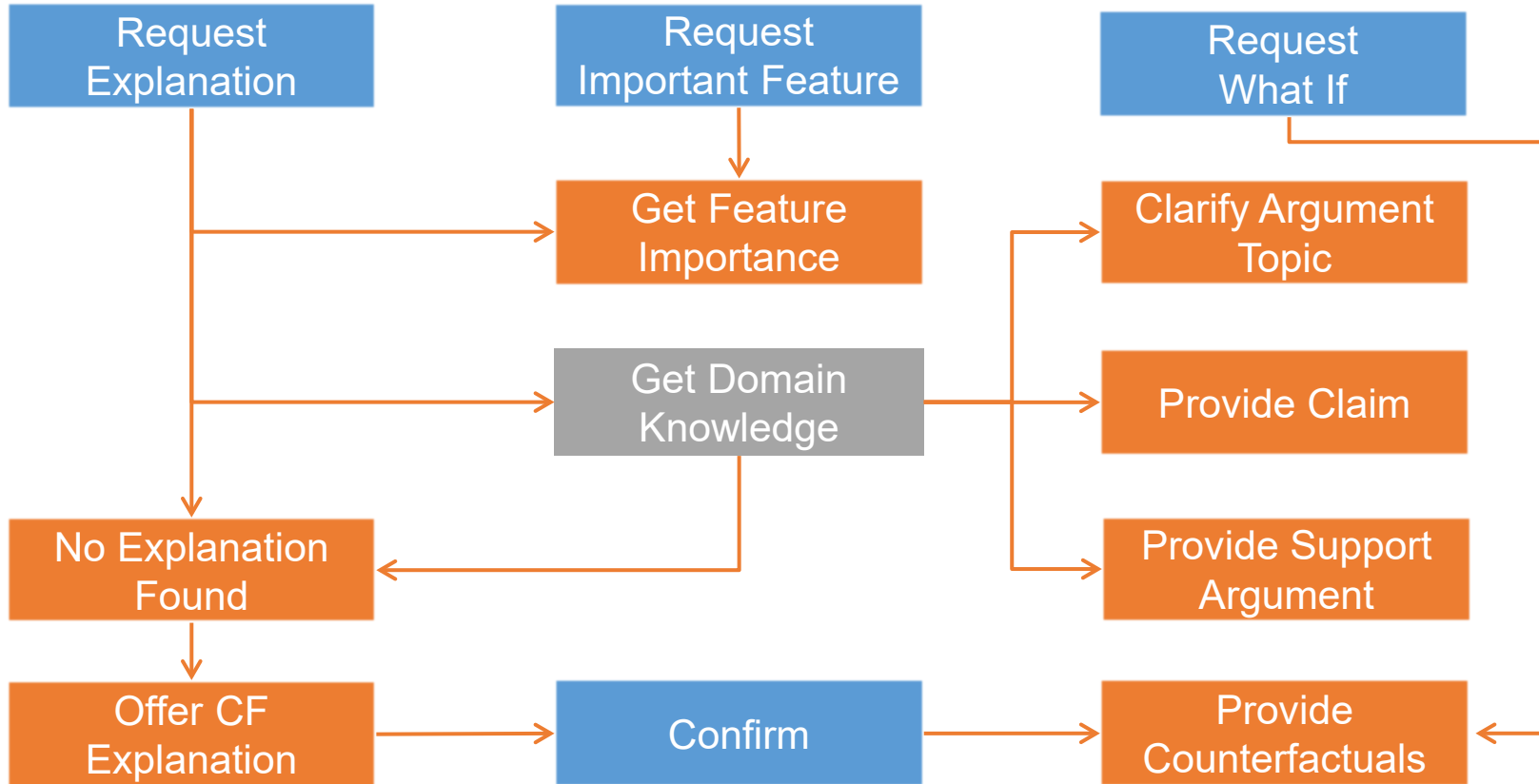
- Two iterations of document processing
  - List of manual gathered domain knowledge documents
  - LLM summarizations for specific questions



# The Impact of Domain Knowledge in Explanatory Dialogues

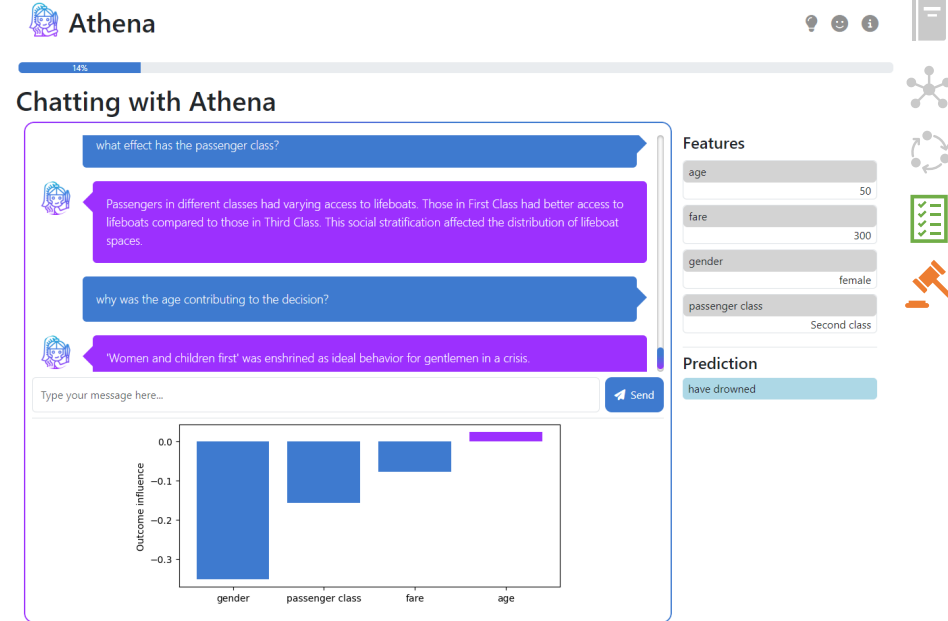


# Explanation Strategy



# Introduction Athena

- Explanatory chatbot for prediction tasks
- Three datasets included (extensible):
  - German credit data („Am I creditworthy?“)
  - Titanic („Would I have survived on the titanic?“)
  - Diabetes Risk Assessment („Am I at risk for diabetes?“)
- Random Forest Classifier
- Three XAI-Methods supported: SHAP, CF, Example



**Athena**

14%

**Chatting with Athena**

what effect has the passenger class?

Passengers in different classes had varying access to lifeboats. Those in First Class had better access to lifeboats compared to those in Third Class. This social stratification affected the distribution of lifeboat spaces.

why was the age contributing to the decision?

"Women and children first" was enshrined as ideal behavior for gentlemen in a crisis.

Type your message here... Send

**Features**

age	50
fare	300
gender	female
passenger class	Second class

**Prediction**

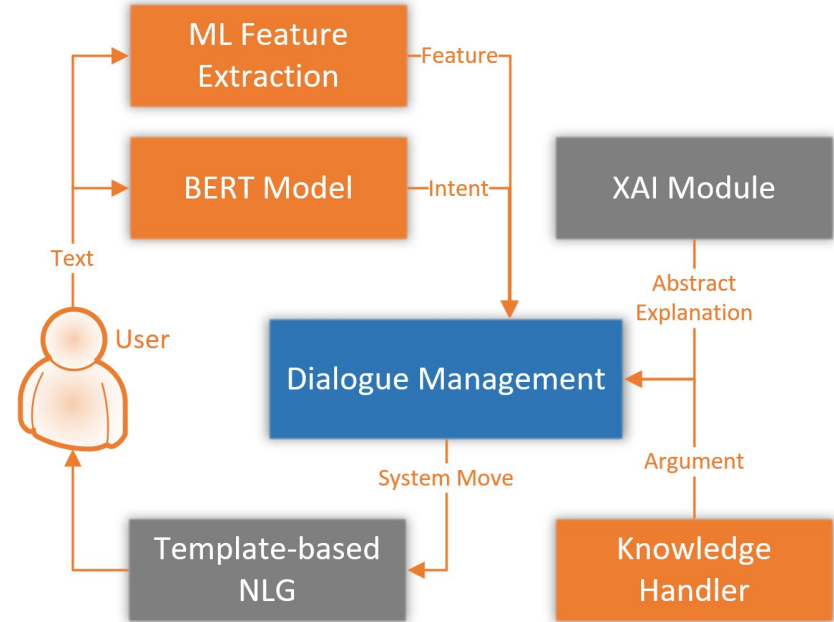
have drowned

**Outcome influence**

Feature	Outcome influence
gender	-0.25
passenger class	-0.15
fare	-0.05
age	0.05

# Athena Architecture

- Explanation intents classified by BERT model (generic); fine-tuned on handcrafted data
- ML Feature Extraction to get use-case specific information
- XAI module with custom implementation of counterfactuals and SHAP-library.
- Rule-based dialogue management
- Knowledge handler for including domain knowledge
- Templates for XAI explanations



# Preliminary Study Setup

- 32 participants in Online Study in 4 groups
- 2 dialogues per participant (with/without domain knowledge)
- True/False AI Setting
- Titanic & Credit Scenario
- Handcrafted and LLM generated arguments with fully manual annotated trees
- Evaluation:
  - 2 XAI related questions (Agreement AI decision)
  - SASSI Questionnaire for dialogue system performance
  - 5-likert Scale
  - Significance measured with Mann-Whitney-U Test



# Preliminary User Study

- Do users understand that the AI is behaving incorrect?  
Does domain knowledge help to detect incorrect AI behavior?
  - Q1: I agree with the decisions made by the system.
  - Q2: The system decisions are plausible.
- Domain knowledge requested by the user: 44%

	AI	No DK avg	DK $\Sigma$	DK avg	DK $\Sigma$	p
Q1	false	2.48	27	2.60	5	0.91
	true	3.69	23	3.89	9	0.87
Q2	false	2.44	27	3.40	5	0.14
	true	3.65	23	4.00	9	0.58

(p is value of Mann-Whitney U Test)





# Preliminary Study Results

## Impact of Domain Knowledge



System likeability is higher (trend true AI, significant false AI)



System appears more robust (false AI)



Reduced cognitive demand (false AI)



System appears more useful (true AI)



# Study Setup

- 80 participants in Online Study
- 4 groups / 2 dialogues per participant (DK & NO DK) / True|False AI Setting
- New scenario diabetes
- New explanation type: example based
- Automatic generated argumentation trees with human in the loop to guarantee quality
- Separation of AI and Dialogue System
- More proactive strategy for Domain Knowledge



# User Study Results

- Domain Knowledge usage increased to 76% (before 44%)
- Q1 I agree with the decisions made by the prediction system.
- Q2 The prediction system's decisions are plausible.

	AI	No DK		DK		$p$	$p^*$
		$\mu$	$\Sigma$	$\mu$	$\Sigma$		
Q1	false	3.22	50	3.57	30	0.18	0.37
	true	3.59	49	3.94	31	0.24	0.48
Q2	false	3.28	50	3.73	30	0.06	0.13
	true	3.69	49	4.00	31	0.13	0.27

Where:

$p$  is value of Mann-Whitney U Test

$p^*$  is the value of Holm-Bonferroni Correction

# User Study Results: Topicwise

- Q1 I agree with the decisions made by the prediction system.
- Q2 The prediction system's decisions are plausible.

	AI	No DK		DK		$p$	$p^*$
		$\mu$	$\Sigma$	$\mu$	$\Sigma$		
Q1	false	3.30	27	4.15	13	<b>0.01</b>	<b>0.03</b>
	true	3.56	25	3.67	15	0.98	0.98
Q2	false	3.33	27	4.08	13	<b>0.03</b>	0.09
	true	3.76	25	3.73	15	0.94	0.94

(a) Credit Scenario

	AI	No DK		DK		$p$	$p^*$
		$\mu$	$\Sigma$	$\mu$	$\Sigma$		
Q1	false	3.13	23	3.12	17	0.93	0.93
	true	3.62	24	4.19	16	0.07	0.22
Q2	false	3.22	23	3.47	17	0.52	0.52
	true	3.62	24	4.25	16	<b>0.03</b>	0.10

(b) Diabetes Scenario

Where:

$p$  is value of Mann-Whitney U Test

$p^*$  is the value of Holm-Bonferroni Correction



# User Study Results – Impact of Domain Knowledge

- longer interaction
- different use of explanation
- small trends in overall dialogue experience
  - more engaging/enjoyable
- challenge to deal with over reliance
- DK is topic dependent

Explanation	No DK available	DK available	<i>p</i>
Counterfactuals	71%	36%	< 0.001
Shapley Values	90%	90%	1.000
Example-based	18%	26%	0.251
New Prediction	21%	24%	0.850
Domain Knowledge	-	76%	< 0.001



# Conclusion

- Transparency is needed for AI
- Conversational XAI enables interactive, personalized experiences
- Argumentation Trees can be used to integrate DK
- DK has impact on understanding and acceptance of AI
- Remaining Challenges
  - Ethical Concerns like over-reliance
  - Truthful sources for DK
  - Domain/Topic dependent



Thank you

“ Knowledge isn't power until it is applied. ”

*Dale Carnegie*



Isabel.Feustel@uni-ulm.de



Find me on LinkedIn



<https://nt.uni-ulm.de/ifeustel>

# Citations

- Antaki, C., & Leudar, I. (1992). Explaining in conversation: Towards an argument model. *European Journal of Social Psychology*, 22(2), 181-194.
- Stab, C., & Gurevych, I. (2014). Annotating argument components and relations in persuasive essays. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 1501–1510.
- Feustel et al. (2024). Enhancing Model Transparency: A Dialogue System Approach to XAI with Domain Knowledge. SigDial 2024
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267, 1-38.
- Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and brain sciences*, 34(2), 57-74.
- Schindler et al. (2025). Automatic Generation of Structured Domain Knowledge for Dialogue-based XAI Systems. IWSDS 2025.