# 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

MARIE CURIE ACTIONS

DIGITAL

# Reading

Slack, D., Krishna, S., Lakkaraju, H., & Singh, S. (2023). Explaining machine learning models with interactive natural language conversations using TalkToModel. arXiv preprint arXiv:2207.04154.

DIGITAL

# Explaining Machine Learning Models with Interactive Natural Language Conversations

Slack, D., Krishna, S., Lakkaraju, H., & Singh, S. (2023). Explaining machine learning models with interactive natural language conversations using TalkToModel. arXiv preprint arXiv:2207.04154.

DIGITAL

Funded by
the European Union

# Motivation

- ML models are widely used in high-stakes domains (healthcare, finance, law).
- Many post-hoc explanation methods exist (LIME, SHAP, counterfactuals, etc.).
- Practitioners struggle because:
  - They don’t know which explanation to use.
  - They don’t know how to interpret the results.
  - They often have follow-up questions that static dashboards cannot easily support.

**Explainability should be a dialogue**, not a one-shot visualization.

# Core idea: TalkToModel

- **TalkToModel** is an **interactive conversational system** that lets users ask questions about a trained ML model in **natural language**, such as:

"Why was this person denied a loan?"

"What features matter most for people over 30?"

"What would need to change to flip this prediction?"

"What kinds of cases does the model get wrong?"

- Instead of clicking through dashboards, users **talk to the model**, and the system responds with explanations grounded in actual model computations.

DIGITAL

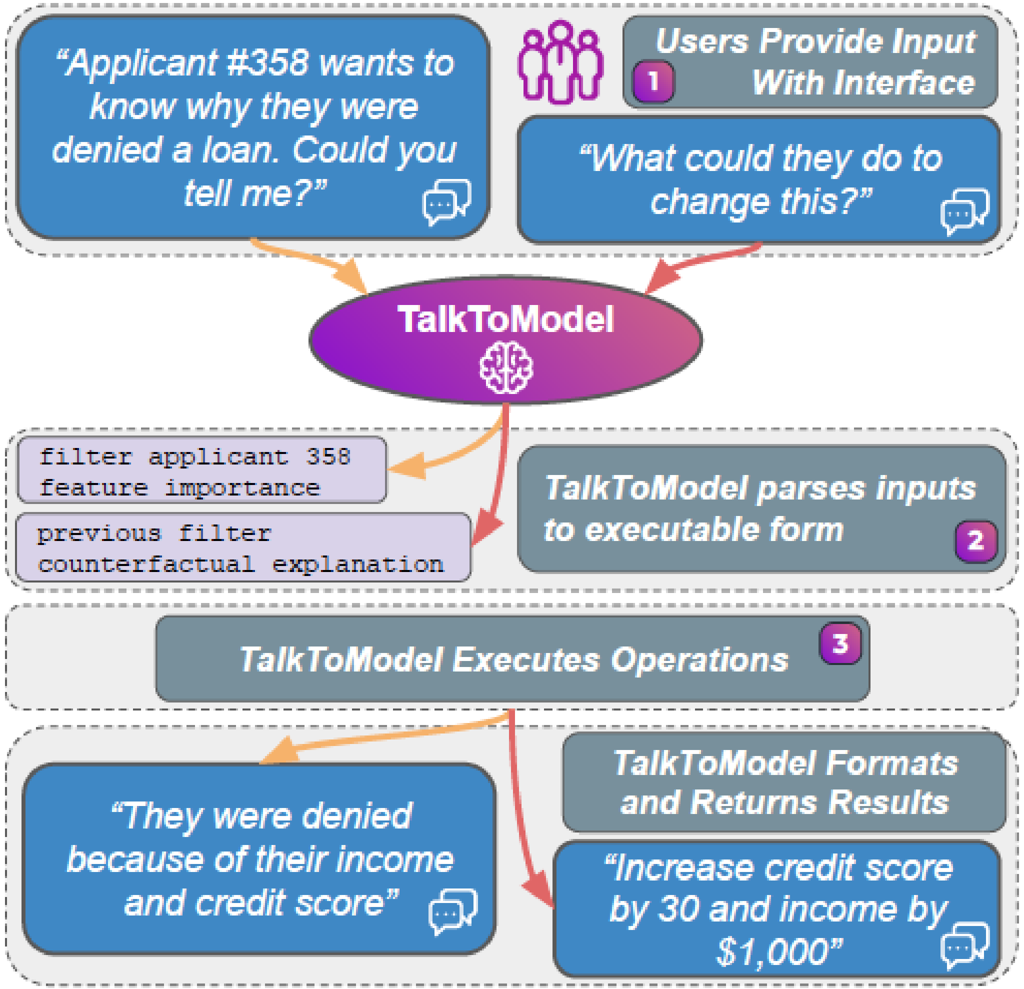


Figure 1: **Overview of TalkToModel:** Instead of writing code, users have conversations with TalkToModel as follows: **(1)** users supply natural language inputs. **(2)** the dialogue engine parses the input into an executable representation. **(3)** the execution engine runs the operations and the dialogue engine uses the results in its response.

DIGITAL

# Component 1: **Dialogue engine (language understanding)**

- User utterances are translated into a **structured, SQL-like program** (a domain-specific language).
- This is done using **large language models (LLMs)** (T5 or GPT-style models).
- The system does *semantic parsing*, not free-text generation:
  - This ensures **accuracy and trustworthiness**
  - Avoids hallucinated explanations

- Example:
- "How important is BMI for patients older than 30?"
- → Parsed into operations like:
- filter(dataset, age > 30)
- explain(dataset, feature_importance)
- important(dataset, BMI)

DIGITAL

# Component 2: **Execution engine (model & explanation computation)**

- Once the program is constructed, the execution engine:
- Runs predictions
- Computes explanations
- Performs filtering, counterfactual analysis, error analysis, etc.
- **Key design choice**: Users do *not* need to choose an explanation method.
- **Automatic explanation selection**
  - The system runs multiple explanation methods (e.g., LIME, SHAP).
  - It selects the **most faithful** explanation using a perturbation-based metric (“fudge score”).
  - If multiple explanations are similar, it prefers the **more stable** one.
- This directly addresses a major real-world pain point: *“Which explainer should I trust?”*

DIGITAL

# Component 3: **Conversation interface**

- Results are returned as **natural language responses**, using templates grounded in the executed operations.

- Conversation history is used:
    - "What if we increased glucose by 10?"
    - "What about for the same group as before?"

- This is handled explicitly using **conversation-aware grammar operations**, not black-box dialogue state tracking.

DIGITAL

Table 1: A conversation about diabetes prediction, demonstrating the breadth of different conversation points the system can discuss.

| User "utterances" & TalkToModel response |
|---|
| "What are the predictions on patients older than 30? Could you also tell me what would happen to these predictions if we increased bmi by 10?" |
| On data with **age greater than 30.0**, the model predicts:<br>➩ **No Diabetes** on 46.8% of data<br>➩ **Diabetes** on 53.2% of data<br>On data with **age greater than 30.0 and bmi is increased by 10.0**, the model predicts:<br>➩ **No Diabetes** on 30.1% of data<br>➩ **Diabetes** on 69.9% of data |
| "So how important is bmi for the model's predictions on the over thirty group?" |
| For the model's predictions on data with **age greater than 30.0**, the **bmi** feature is ranked on average **2.86** (95% CI [2.685, 3.035]). Here, rank 1 is the most important feature (out of 8 features). Overall, **bmi** is **highly important** for the model's predictions. |
| "Ah, so which two features are mostly more important than bmi for these people?" |
| For the models predictions on data **where age is greater than 30.0**, the **top 2** most important features are:<br>1. **glucose**<br>2. **pedigree_function** |
| "What types of patients is the model typically predicting incorrect here"? |
| For data with **age greater than 30.0**, the model typically predict predicts incorrect:<br>➩ if (**bmi > 26.95**) and (**glucose ≤ 125.0**) and (**pedigree_function** then the model is incorrect on *62.3%* of *53* samples<br>➩ (**bmi > 26.95**) and (**glucose > 125.0**) and (**age > 57.5**) **<= 0.943**) then the model is incorrect on *62.5%* of *8* samples<br>➩ (**bmi ≤ 26.95**) and (**glucose > 180.5**) then the model is incorrect on *100.0%* of *2* samples.<br>Want to take a closer look at these rules? |

DIGITAL

# What kinds of questions can it answer?

- The system supports a **wide range of XAI-relevant queries**, including:
- **Model behavior**
  - Predictions and probabilities
  - Performance metrics
  - Error patterns
- **Explanations**
  - Feature importance (global & local)
  - Counterfactual explanations ("what needs to change?")
  - Feature interactions
- **Data exploration**
  - Filtering subgroups
  - Summary statistics
  - Distributional questions
- The authors show their grammar covers **30 out of 31** core XAI questions identified in prior human-centered XAI research.

DIGITAL

# Evaluation: Language understanding accuracy

- They evaluate how well TalkToModel translates natural language into correct programs:
- Datasets: Diabetes, German Credit, COMPAS
- Metric: **Exact match parsing accuracy**
- Result:
  - Fine-tuned **T5 models perform best**
  - Strong performance even on *compositional* (previously unseen) queries
  - Much better than few-shot GPT-J prompting
- This shows the system can reliably understand user intent.

DIGITAL

Table 3: Exact Match Parsing Accuracy (%) for the 3 gold datasets, on the IID and Compositional splits, as well as Overall. The fine-tuned T5 models perform significantly better than few-shot GPT-J, and T5 Large performed the best. These results demonstrate that TalkToModel can understand user intentions with a high degree of accuracy using the T5 models.

| | German | | | Compas | | | Diabetes | | |
|---|---|---|---|---|---|---|---|---|---|
| | IID | Comp. | Overall | IID | Comp. | Overall | IID | Comp. | Overall |
| Nearest Neighbors | 26.2 | 0.0 | 16.5 | 27.4 | 0.0 | 21.9 | 10.9 | 0.0 | 8.4 |
| GPT-Neo 1.3B | | | | | | | | | |
| 10-SHOT | 41.3 | 4.1 | 27.5 | 35.9 | 0.0 | 28.8 | 40.1 | 7.0 | 32.6 |
| 20-SHOT | 39.7 | 0.0 | 25.0 | 39.3 | 0.0 | 31.5 | 42.9 | 2.3 | 33.7 |
| 30-SHOT | 42.9 | 0.0 | 27.0 | 39.3 | 0.0 | 31.5 | 41.5 | 4.7 | 33.2 |
| GPT-Neo 2.7B | | | | | | | | | |
| 5-SHOT | 38.1 | 4.1 | 25.5 | 35.9 | 3.4 | 29.5 | 46.9 | 7.0 | 37.9 |
| 10-SHOT | 38.1 | 6.8 | 26.5 | 40.2 | 3.4 | 32.9 | 40.8 | 9.3 | 33.7 |
| 20-SHOT | 39.7 | 0.0 | 25.0 | 39.3 | 0.0 | 31.5 | 42.9 | 2.3 | 33.7 |
| GPT-J 6B | | | | | | | | | |
| 5-SHOT | 51.6 | 14.9 | 38.0 | 51.3 | 6.9 | 42.5 | 55.8 | 7.0 | 44.7 |
| 10-SHOT | 57.9 | 9.5 | 40.0 | 49.6 | 3.4 | 40.4 | 53.7 | 9.3 | 43.7 |
| T5 | | | | | | | | | |
| SMALL | 61.1 | 32.4 | 50.5 | 71.8 | 10.3 | 59.6 | 77.6 | 30.2 | 66.8 |
| BASE | 68.3 | **48.6** | 61.0 | 65.0 | 10.3 | 54.1 | **84.4** | 34.9 | 73.2 |
| LARGE | **74.6** | 44.6 | **63.5** | **76.9** | **24.1** | **66.4** | **84.4** | **51.2** | **76.8** |

# Evaluation: Human user study

- They compare TalkToModel against a popular explainability dashboard.
- **Participants**:
  - 45 healthcare workers (minimal ML background)
  - 13 ML professionals
- **Findings**:
  - Users were:
    - **Faster**
    - **More accurate**
    - **More confident**
  - Healthcare workers strongly preferred TalkToModel
  - Even ML experts answered **more questions correctly** with TalkToModel
- This empirically supports the claim that **conversation is a better interface for explainability**.

DIGITAL

Table 4: User study results: % of respondents that agree (> Neutral Likert score) TalkToModel is better than the dashboard in the 4 comparison questions. A significant portion of respondents agreed TalkToModel is better than the dashboard in all the categories except Grad. students and "Likeliness To Use". Still, a majority agreed TalkToModel was superior in this case.

| | % Agree TalkToModel Better | |
|---|---|---|
| Comparison | Health Care Workers | ML Grad. Students |
| Easiness | 82.2 | 84.6 |
| Confidence | 77.7 | 69.2 |
| Speed | 84.4 | 84.6 |
| Likeliness To Use | 73.3 | 53.8 |

DIGITAL

# Limitations

- **Limitations and future work (as acknowledged)**
- Currently focused on **tabular data**
- Explanations are mostly text-based (limited visualization)
- Expert users may still want raw data access
- Future directions include:
  - Richer visual grounding
  - Deployment in real operational settings
  - More adaptive explanation generation

DIGITAL

DIGITAL

Funded by
the European Union

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.

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*