



UNIVERSITY  
OF AMSTERDAM



# Actionable Interpretability for Churn Classification: A Text Bottleneck Model Case Study at a Major Telecom Provider

Paper ID: #235

***Adrian Sauter<sup>1,\*</sup>, Vera Neplenbroek<sup>1</sup>, Giorgos Vlassopoulos<sup>2</sup>, Gianluigi Bardelloni<sup>2</sup>***

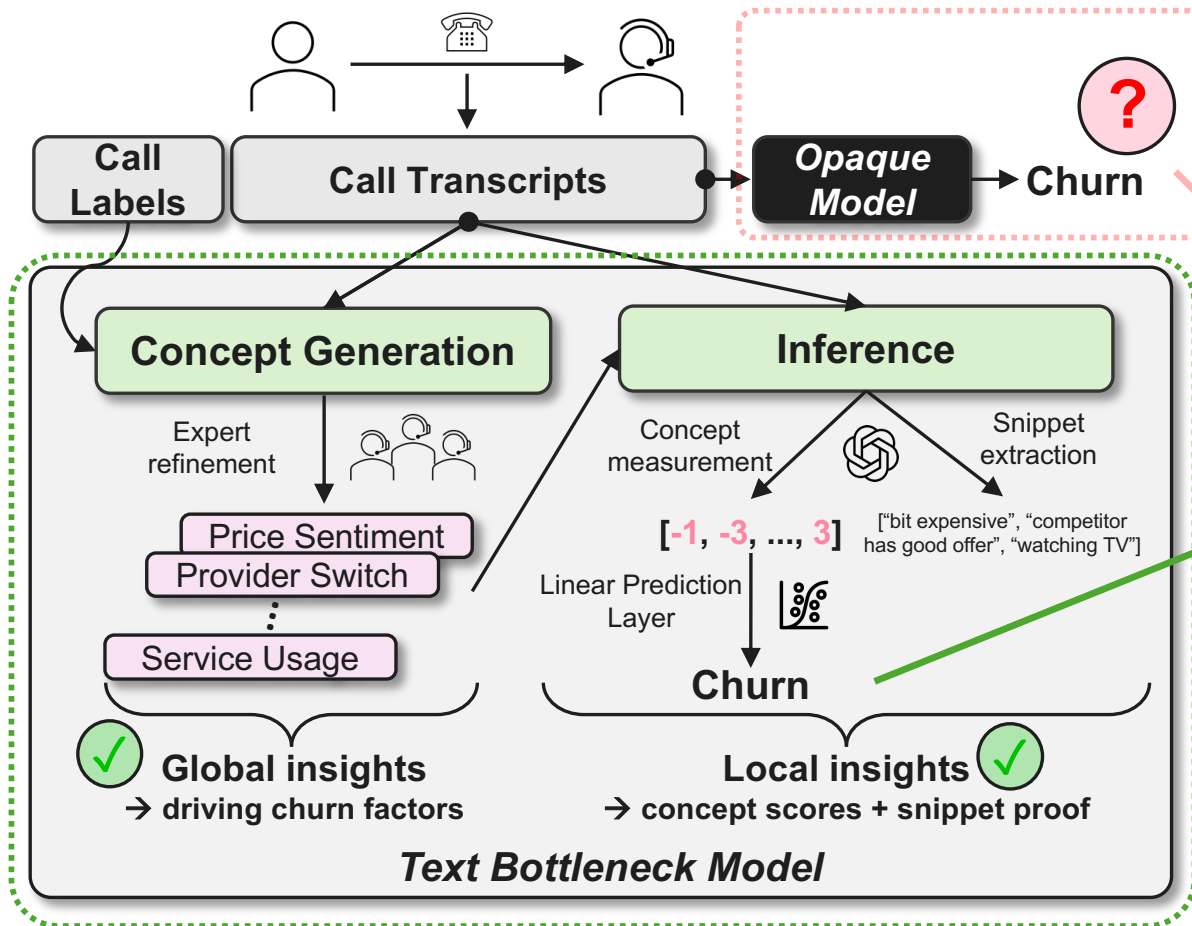
*<sup>1</sup>University of Amsterdam, <sup>2</sup>KPN, \*now at Helmholtz Munich*

Correspondence:

*adrian.sauter@helmholtz-munich.de*



# Background & Motivation



## Context

- Managing customer churn is vital for subscription-based businesses

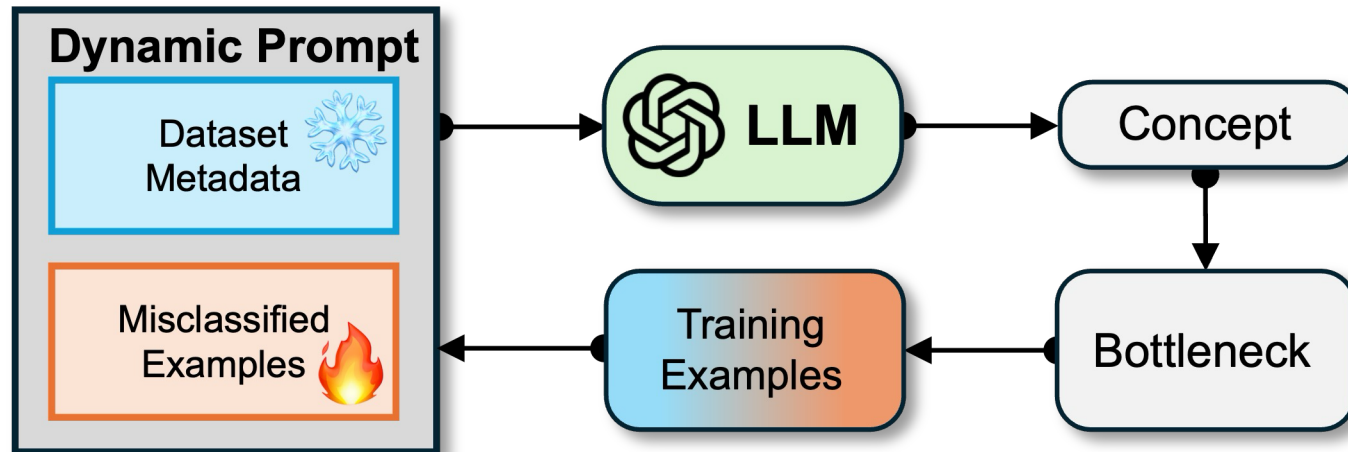
## Problem

- Current models classify churn but don't explain it — leaving retention teams without guidance

## Our solution

Text Bottleneck Model (TBM) [1]: interpretable concept layer → actionable retention guidance

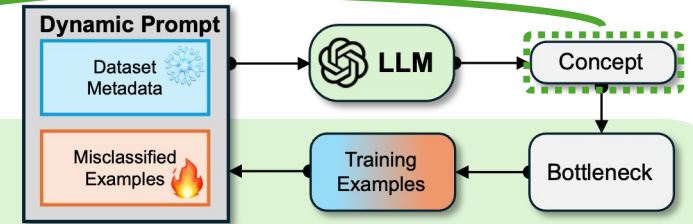
## Text Bottleneck Model – Concept Generation



LLM iteratively proposes concepts targeting previously misclassified examples

# Text Bottleneck Model – Concept Example

## Relocation Mention



- **Description:** Relocation Mention captures whether the customer mentions moving to a new home, and whether this move is associated with continuing or canceling [TelCo] service.
- **Question:** Does the customer mention moving houses, and do they express an intent to cancel or continue their [TelCo] service as a result?
- **Response Guide:**
  - *Moving and Canceling Service (-3):* The customer clearly states that they are moving and plan to cancel their [TelCo] service or taking a subscription from a different provider at the new address.
  - *Possible Relocation Mentioned (-1):* There is an indirect mention or hint of a move, without confirmation or clarity on service continuation or cancellation.
  - *No Mention of Relocation (0):* There is no mention of moving houses or any change in living situation during the call.
  - *Moving and Continuing Service (3):* The customer mentions moving but also expresses an intent to keep using [TelCo], such as arranging a transfer of service.

# Text Bottleneck Model – Concept Measurement

Given a transcript, an LLM is prompted with each concept definition

<p><b>Concept: Relocation Mention</b></p> <p>[agent]: Good afternoon, welcome to [TelCo], how can I help you? [customer]: I'm <b>moving soon and would like to cancel my subscription</b>, because I'm not using it anymore.</p>
<p><i>Response: Moving and canceling service Concept Score: -3</i></p> <p><i>Explanation: One snippet for <b>Moving and canceling service</b>. The customer mentions that they are moving and that they want to terminate the service.</i></p>

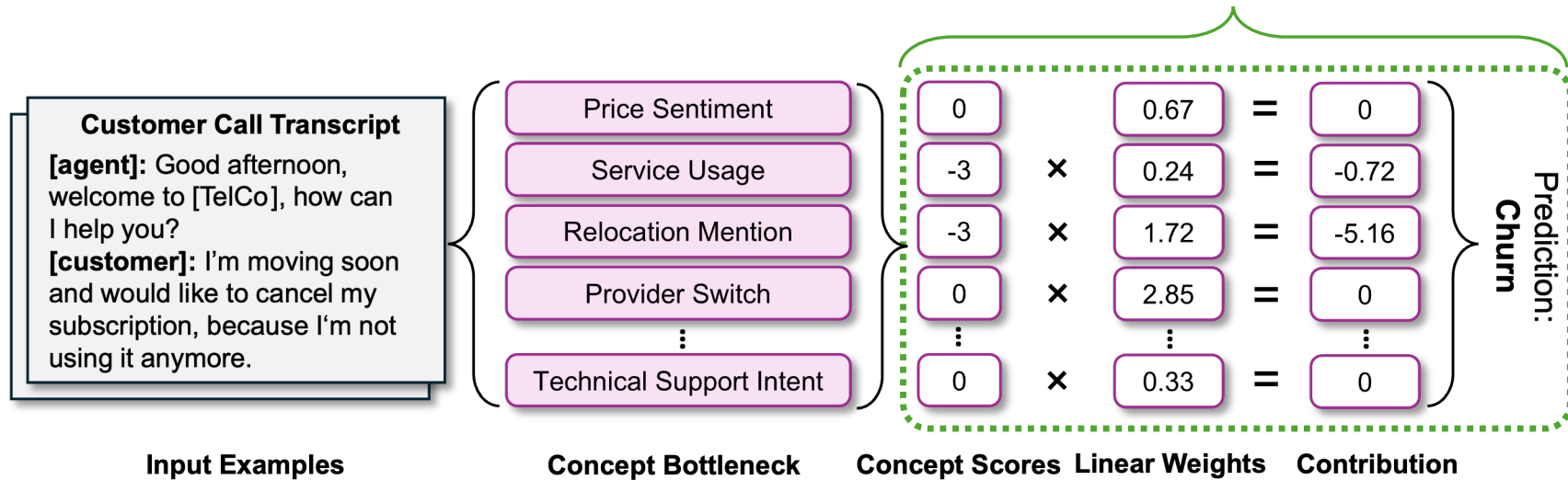
Responses are mapped to numerical scores

LLM extracts snippets + thought section for final measurement

→ **Local interpretability:** Extracted snippets and concept scores offer local explanations

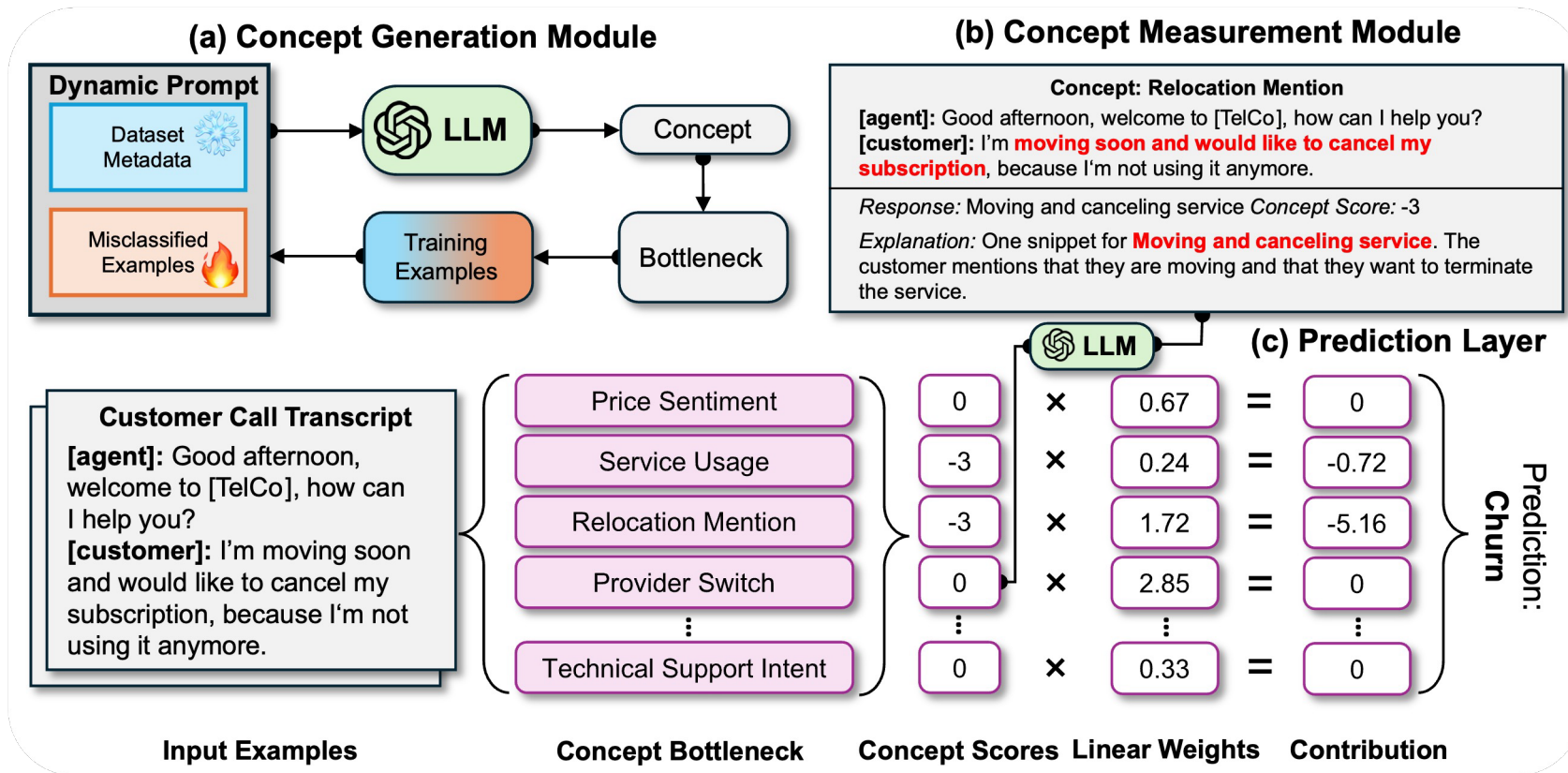
# Text Bottleneck Model – Prediction Layer

White-box classifier takes vector of concept measurements and predicts target label



→ **Global interpretability:** Learned weights indicate concept importance across dataset

# Text Bottleneck Model – Overview



# Experimental Setup – Dataset & Models

## Dataset

- Expert-annotated dataset in Dutch
- Artificially balanced **train set** (100 Churn, 100 No Churn samples)
- „Natural“ **test set**: 28 Churn, 172 No Churn samples



## Models

- TBM: **gpt-4-turbo-128k**, gpt-4o-mini
- Baseline (**black-box**): Domain-adapted BERTje [2], gpt-4-turbo-128k, gpt-4o-mini



## Results – Churn Prediction

Framework	Prediction Model	F1-Score	Cohen's $\kappa$	AUPRC
Black-Box Production Model	BERTje	0.9081	0.8429	–
Black-Box GPT-4	gpt-4-turbo-128k	0.9485	0.8996	–
TBM	Logistic Regression (no reg.)	0.9198	0.8655	0.9254
	Logistic Regression (L1-reg.)	0.9184	0.8621	0.9230
	Logistic + Interactions (no reg.)	0.8780	0.7984	0.8812
	Logistic + Interactions (L1-reg.)	0.8615	0.7752	0.8645
	Decision Tree	0.8835	0.8091	0.8522

**Finding 1:** TBM with 7 expert-refined concepts performs competitively with blackbox models

**Finding 2:** Expert-refined concepts outperform automatic concept generation (max. F1-score: 0.7325)

# Results – Final Set of Concepts

*stronger churn signal → weaker (log. Regression coefficients)*

**-2.85**

## **Provider Switch**

Customer mentions switching to or from a competitor

**-1.77**

## **Service Modification Intent**

Customer inquires about downgrading or upgrading service

**-1.72**

## **Relocation Mention**

Customer indicates they are moving and may cancel

**-0.67**

## **Price Sentiment**

Customer mentions satisfaction/dissatisfaction with pricing

**-0.61**

## **Subscription Flexibility Concern**

Customer mentions contract flexibility

**-0.33**

## **Technical Support Intent**

Customer mentions tech issues and whether they seek help

**-0.24**

## **Service Usage**

Customer mentions whether they have stopped using the service

# Results – Qualitative Error Analysis

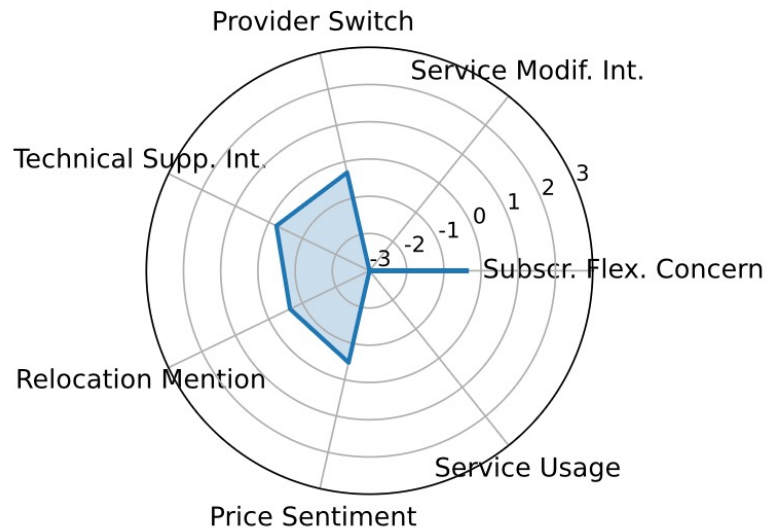
	Churn intent present	Churn intent absent
Concepts activated	<p><b>Correct prediction: Churn ✓</b></p> <p><i>"I'm moving and want to cancel my subscription"</i> → Relocation Mention: -3 → Churn</p>	<p><b>False positive: Churn ✗</b></p> <p><i>"The price is too high and the service is inflexible"</i> → Price Sentiment + Subscription Flexibility activate, but customer stays</p>
No concepts activated	<p><b>False negative: No Churn ✗</b></p> <p><i>Concept scores don't cross threshold: "I'm a bit unhappy with the price and the contract flexibility and want to switch"</i> → score too low</p> <p><i>No concept triggered: "I just want to stop" → no reason given</i></p>	<p><b>Correct prediction: No Churn ✓</b></p> <p><i>"I'd like to upgrade my plan"</i> → no churn concepts activate → No Churn</p>

# Results – Identifying Call Profiles

K-means in concept space to identify profiles (K=4)

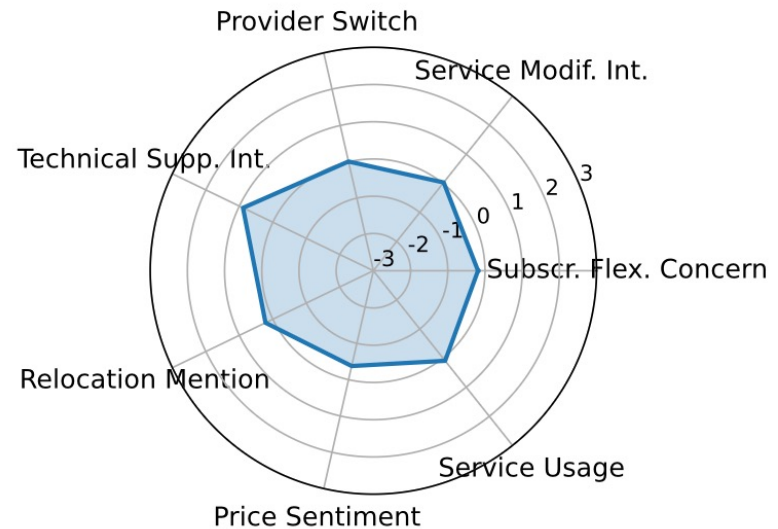
Cluster 2 Profile (108 samples)

Actual Churn Rate: 98.1% | Predicted Churn Rate: 99.1%



Cluster 4 Profile (128 samples)

Actual Churn Rate: 15.6% | Predicted Churn Rate: 7.8%



**Finding 3: Concept space contains business-relevant information that is easily accessible**

## Results – Additional Results

---

### Finding 4

0.943

avg. triple agreement

→ **Concept measurements are consistent**

---

Repeated concept scoring produces stable results across runs

### Finding 5

88%

direct transcript matches

→ **Snippets are faithful and sufficient**

---

Independent LLM can predict concept scores from snippets alone (~90% accuracy) — no full transcript needed

### Finding 6

84.49%

F1 without chain-of-thought

→ **Chain-of-thought improves performance (91.98%)**

---

Removing the „thoughts“ section from the prompt measurably reduces F1-score

# Interactive Dashboard

## 🔍 Concept Explorer for Call Transcripts

### 📘 About This Webapp

This dashboard uses **Concept Bottleneck Models** to make AI decisions transparent — breaking down customer service calls into key human-understandable concepts.

**How it works:** Enter a contact ID to fetch a transcript. An LLM scores key concepts and highlights the exact parts of the conversation behind each score. You can also learn more about the individual concepts by clicking on 'Show Concept Details'.

**Churn Prediction:** Once analyzed, you'll see two churn classifications: one from the production TECLA model, and one based on concept-driven reasoning.

### 📞 Search by Contact ID

✅ Processing Complete!

Sample fetched and labeled successfully!



TECLA Classification: Churn

Concept Bottleneck Model Classification: Churn

### Concept Progress:

Price Sentiment



Provider Switch



Relocation Mention



Service Modification Intent



Service Usage



Subscription Flexibility Concern



Technical Support Intent



# Interactive Dashboard

🔍 Concept Measurement & Snippet Highlighting

Provider Switch ✕

**Model Measurement:**

**competitor mentioned directly**

**Model Thoughts:**

The customer explicitly mentions Ziggo and their offering as a reason for considering a switch away from KPN. They also state their intent to switch specifically to Ziggo, making this a clear case of 'competitor mentioned directly'. There is no indication of a switch to KPN or a lack of provider switch discussion.

Highlight Specific Options:

- competitor mentioned directly
- competitor implied
- no provider switch mentioned
- switch to KPN

[Hide Concept Details](#)

📄 Transcript Viewer

[Agent]: Good afternoon, how can I help you?

[Customer]: Hi, I'm going to move within a couple of weeks and therefore I do not want to prolong my subscription. I want to check my options with you.

[Agent]: Of course, may I ask you why you don't want to prolong your subscription?

[Customer]: Actually I've been very happy so far with the prive and services of [Telco]. So it does not depend on those.

[Agent]: Nice to hear, what then makes you want to terminate your subscription?

[Customer]: With this relocation I prefer to have flexibility. I saw that [Company] has a move offer where I can get a shorter contract. With [TelCo], I'd be committed for a whole year and that does not with with my situation. Other providers are not so demanding.

## Provider Switch

Provider Switch captures whether the customer discusses switching between telecommunication providers, either away fromTelCo to a competitor, or from a competitor toTelCo. It includes both explicit and implicit mentions of such switches.

### LLM Prompt Question:

Does the customer mention switching away from TelCo to another provider, or switching from another provider to TelCo?

## Response Guide:

- competitor mentioned directly**: The customer explicitly states they plan to or have recently switched away fromTelCo to a specific competitor, naming that provider (e.g., [company1], [company2], [company3]). Use this only if the customer is currently leavingTelCo or expresses clear intent to do so. Examples: 'I switch to [company1]', 'At [company2] I get more for less', 'I just signed up at [company3].'
- competitor implied**: The customer expresses an intent to leaveTelCo or mentions receiving a better offer but does not name the competitor. Use this only if the customer is actively considering or planning to leave TelCo. Example: 'I got a better offer from another provider', 'I am thinking about changing, this is too expensive.'

- no provider switch mentioned**: The customer does not mention any switch to or fromTelCo. The transcript is not related to any provider switch (either explicitly or implicitly).
- switch to TelCo**: The customer explicitly states or clearly implies that they have recently switched from a competitor toTelCo. This includes both explicit mentions or clear implications of switching in favor ofTelCo. Examples: 'I was at [company1],' or 'I am happy that I switched to TelCo.'

# Takeaways

---

## Interpretability without compromise

---

TBMs match black-box performance while providing both global and local interpretability into why customers churn

## Concepts unlock actionable insights

---

Concept space enables customer trend analysis, what-if scenarios, and targeted retention strategies

## A general-purpose framework

---

TBMs apply to any classification or regression problem

Thank you for your attention! :)



Paper



Adrian Sauter

## References

---

[1] Josh Magnus Ludan, Qing Lyu, Yue Yang, Liam Dugan, Mark Yatskar, and Chris Callison-Burch. 2023. Interpretable-by-design text classification with iteratively generated concept bottleneck. *arXiv preprint arXiv:2310.19660*.

[2] Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert model. *ArXiv, abs/1912.09582*.