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

Adrian Sauter[1,*], Vera Neplenbroek[1], Giorgos Vlassopoulos[2], Gianluigi Bardelloni[2]

[1]University of Amsterdam, [2]KPN, [*]now at Helmholtz Munich

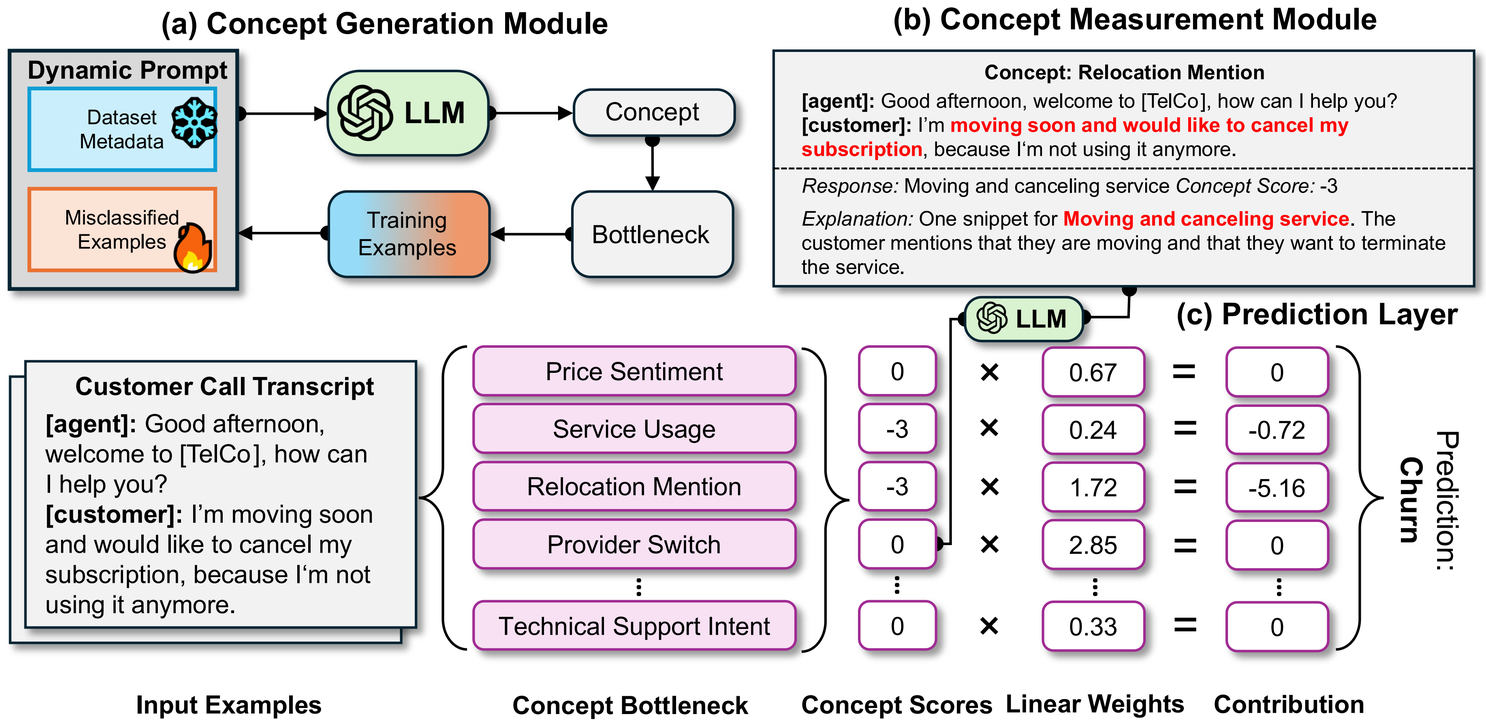
Motivation

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



Example Concept

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.

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

Predictive Performance

Framework	Prediction Model	F1-Score	Cohen's κ	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)

Additional Results

Finding 4
0.943
avg. triple agreement

→ Concept measurements are consistent

Repeated concept scoring produces stable results across runs

Finding 5
84.49%
F1 without chain-of-thought

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

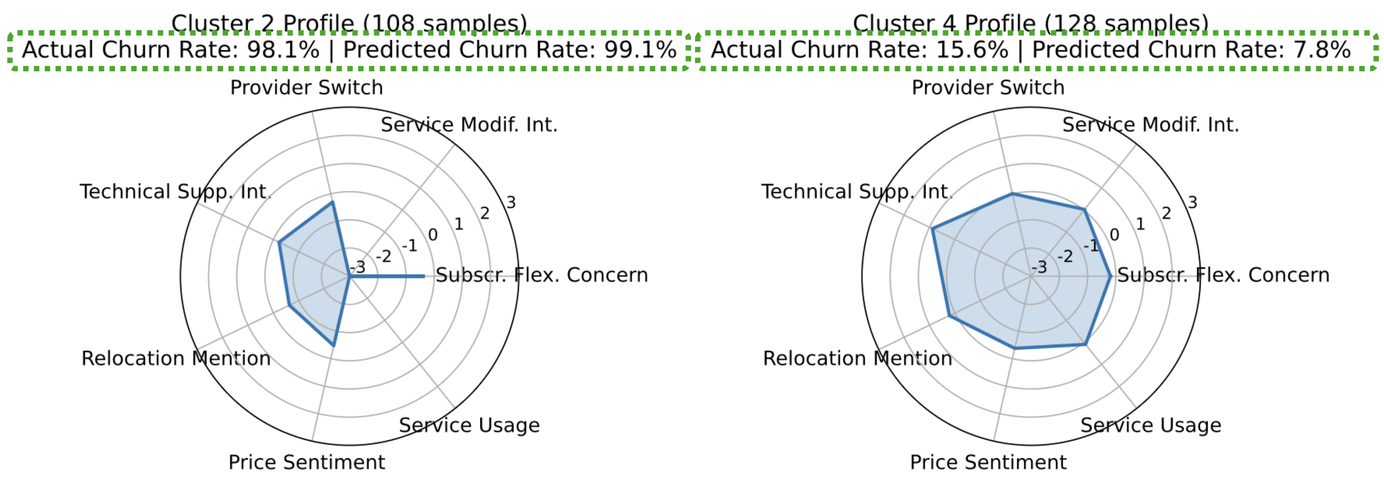
Removing the „thoughts" section from the prompt measurably reduces F1-score

Finding 6
88%
direct transcript matches

→ Snippets are faithful and sufficient

Independent LLM can predict concept scores from snippets alone — no full transcript needed

Identifying Call Profiles



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

Paper

Adrian Sauter

ACL 2026 SAN DIEGO JULY 2–7

[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*.