Social Computing Research Group
TUM School of Computation, Information and Technology
Technical University of Munich

Leveraging LLM Pipelines to Generate Product Comparison Data

Khalil Chikhaoui, Jaeyeop Chung, Umut Ekin Gezer, advisor: Dr. Gerhard Johann Hagerer

Practical Goals

1

- **Dataset Creation:** Develop a comprehensive dataset of insurance products and their comparisons extracted from company websites.
- **Automation:** Automate data collection, classification, data quality improvement and comparison processes to streamline analysis and reduce manual effort.
- **Market Insights:** Provide real-time insights into the insurance market by comparing products across companies.

Technical Goals

2

- **Modular Pipelines:** Showcase the effectiveness of modular pipelines using Kedro.
- **LLM Prompting:** Demonstrate optimal setups for improving LLM prompt performance.
- **NLP Integration:** Leverage NLP and LLMs for structured analysis of insurance products.
- **LLM-as-a-Judge:** Explore the LLM-as-a-judge methodology to evaluate and improve results.
- **Prompt Optimization Pipelines:** Develop data pipelines specifically tailored for reproducible prompt optimization and evaluation

Pipeline Implementation in Kedro

3

❖ The entire data processing workflow was implemented within **Kedro**, leveraging its **modular design**, **data management capabilities**, and **scalability** to create a structured, maintainable pipeline. The pipeline consists of **six main nodes**, each responsible for a specific task:

➢ **1. Data Crawling:**
- **Objective:** Collect website data from **8 German insurance companies:**
  - ARAG, AXA, Barmenia, DKV, ERGO, Generali, HUK-COBURG and SDK.
- **Depth:** Crawled up to **200 pages** per company to ensure thorough coverage of their online content.
- **Tool:** **BeautifulSoup** was used for parsing and extracting raw HTML data from crawled websites.
- **Output:** A dataset of HTML pages stored in the Kedro data catalog for further processing.

➢ **2. Filtering Relevant Pages:**
- **Objective:** Identify pages specifically related to **insurance products.**
- **Method** Filtered pages based on **URL and file name** relevance using **LLM-based prompts**.
- **Output:** A set of filtered pages relevant to **insurance products**.

➢ **3. HTML to Markdown Conversion:**
- **Objective:** Simplify webpage content for easier analysis.
- **Method:** Converted filtered HTML pages into **markdown format** while preserving essential information.
- **Output:** **Clean and structured markdown** files ready for feature extraction.

➢ **4. Feature Extraction:**
- **Objective:** Extract insurance **product names and their relevant features** from markdown files.
- **Method:**
  - Used LLM-based extraction to identify: **Product names** and **Features**.
  - Stored extracted details in a structured format for classification.
- **Output:** A dataset containing **detailed product information.**

➢ **5. Product Classification:**
- **Objective:** Categorize insurance products based on the **Wikipedia Versicherungsart list** (Rentenversicherung, Krankenversicherung, Haftpflichtversicherung…).
- **Method:**
  - Used reference list categories to guide LLM-based classification.
  - Ensured consistency and alignment with standard insurance categories.
- **Output:** A **classified list of insurance products** grouped by **type**.

➢ **6. Product Comparison:**
- **Objective:** Compare products across companies and within the same insurance type.
- **Method:**
  - Product Matching: Used **SentenceTransformer embeddings** with cosine similarity (**>0.7**) to detect highly similar products within the same category.
  - LLM-Based Comparison: Used **GPT-4o** to generate structured Markdown tables, comparing shared features, pricing, and coverage while highlighting key differences.
- **Output:** A **detailed comparison tables** summarizing product attributes.

Generated Comparison Table

Feature	SDK	ARAG
Produktname	Auslandskrankenversicherung der SDK	Auslandsreisekrankenversicherung
Ärztliche Behandlung	Ambulante & ärztliche Behandlung, Arznei-, Verbands- und Hilfsmittel	Ärztliche Heilbehandlung, Arznei-, Verband-, Heilmittel und Röntgendiagnostik
Krankenhausaufenthalt	Behandlung und Unterbringung im Krankenhaus	Stationäre Heilbehandlung, Unterbringung sowie Operationen und Operationsnebenkosten
Zahnbehandlung	Zahnbehandlung und Zahnersatz	Schmerzstillende Zahnbehandlung, notwendige Füllungen, Reparaturen von Zahnersatz
Krankenrücktransport	Medizinisch sinnvoll und vertretbar	Medizinisch sinnvoller und vertretbarer Krankenrücktransport
Notrufservice	24/7 Notrufservice unter +49 711 99522001	24/7 ARAG-Notruf
Familienversicherung	Ehepartner oder Lebensgefährte, Kinder bis 18 Jahre	Ehepartner, Lebensgefährten und Kinder bis zum 18. Lebensjahr
Corona-Abdeckung	Kostenübernahme bei Covid-19-Erkrankung	Behandlungskosten bei Erkrankung durch das Coronavirus
Im Todesfall	Bestattung und Überführung	Überführung ins Heimatland oder Bestattung im Ausland bis 10.250 €
Besonderheiten	Ergänzung zur gesetzlichen Krankenkasse, Schutz der Beitragsrückerstattung	Keine Selbstbeteiligung bei privater Krankenversicherung

Experiments

4

❖ **1. Classification Experiment: Filtering Relevant Pages**

➢ **Objective:** Compare two prompts for classifying product vs. non-product pages, using a self-annotated ground truth data.
- **Basic Prompt (filename-based):** Identifies **52.7%** of pages as **product-related** with **90% accuracy** and **89% precision**, making it suitable for minimizing false positives and ensuring clean downstream analysis.
- **Advanced Prompt (system & human messages):** Identifies **63.1%** of pages as **product-related** with **97% recall**, effectively capturing more relevant pages for competitive intelligence, though with a higher false positive rate.

	Accuracy	Precision	Recall	F1 Score
Classification with Basic Prompt	**0,90**	**0,89**	0,92	**0,91**
Classification with Advanced Prompt	0,85	0,78	**0,97**	0,86

❖ **2. Classification Experiment: Product Classification**

➢ **Objective:** Compare two prompting strategies for categorizing insurance products.
- **Basic Prompt :** Uses only the product name and brief details for classification.
- **Advanced Prompt (system & human messages):** Incorporates structured attributes such as **coverage**, **price**, and **exclusions** for greater classification accuracy.

➢ **Results:**The Advanced Prompt **labels 90.83% of products**(vs. **84.97%** for the Basic Prompt), with the additional insurance products mostly in Health and Liability categories.

❖ **3. Product Comparison Experiment: Comparison Table Generation**

➢ **Objective:** Evaluate the impact of prompts in the pipeline on product comparison quality.

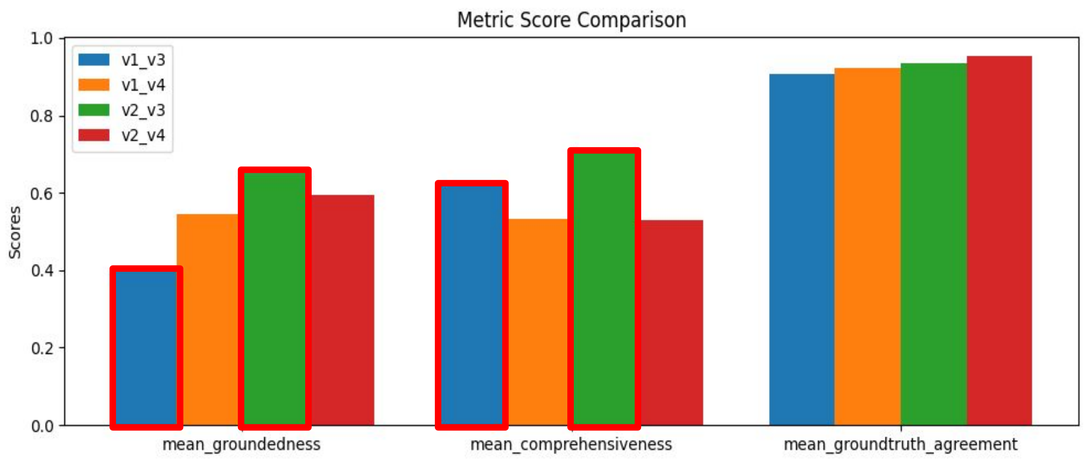
➢ **Method**

Used different prompts(v3, v4) to generate comparison tables from the classified data using the basic prompt(v1) and advanced prompt(v2) in *2. Classification Experiment.*
- **Basic Prompt(v3)** with a **simple command** to generate comparison tables.
- **Advanced Prompt(v4)** with instructions and **structured features to compare** such as coverage, price, and exclusions similar to v2.

➢ **Metrics**

Evaluated using **LLM-as-a-Judge** with the **TruLens** package (score scale: 0~1).
- **Groundedness:** Measures **traceability** to the input source, with lower scores indicating weaker traceability.
- **Comprehensiveness:** Assesses if the generated output contains **key information** from the input source. The more key information, the higher the score.
- **Groundtruth Agreement:** Evaluates how **resemblant** the generated texts are to the original texts. The more resemblant, the higher the score.

➢ **Results**
- **The prompt improvement at the classification step(v1→v2, v3)** showed **the greatest increase in groundedness by 0.25(0.40→0.65)** with **higher comprehensiveness(0.65→0.71);** the classification prompt(v2) notably improved the quality of the product comparisons with more source-based and product information.



	v1_v3_result_tables	v1_v4_result_tables	v2_v3_result_tables	v2_v4_result_tables
Company Count	**8**	**8**	**8**	**8**
Product Count	**532**	**532**	**800**	**800**
Category Count	**23**	**23**	**24**	**24**

❖ **Conclusion**

Higher data quality from improvements in classification prompt(v1→v2) allowed the LLM to extract more relevant features from the source data, as reflected in the notable increase in groundedness and comprehensiveness, leading to more accurate product comparisons.

Limitations

- **Unstandardized Features**: Insurance product attributes (e.g., coverage, price) vary widely, making consistent comparisons difficult.
- **Limited Feature Availability:** Many websites omit or only partially provide key details, reducing classification accuracy.
- **Restricted Category Scope:** Some niche or emerging insurance types may not fit neatly into the chosen categories.

Future Work

5

- **Enhanced Metrics:** Develop additional metrics to further refine and evaluate the prompt enhancement process.
- **Multi-Agent Systems:** Implement multi-agent frameworks to automate prompt enhancement for specific insurance categories.
- **Broader Scope:** Expand the scope of data collection to include companies from other regions (e.g., the US) to generalize prompt performance across diverse markets.