Intelligent Data Integration from Web Sources

Excel@FIT 2026

Author: Bc. Rudolf Jurišica

Supervisor: doc. Ing. Radek Burget Ph.D.

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Problem? A vast amount of inconsistent, unstructured and overlapping data is distributed across multiple places.

Solution? An execution pipeline with a web application that aggregates, extracts, normalizes and extends data, integrating LLMs for natural language search over the unified dataset.

1. Data retrieval

Bazos.cz

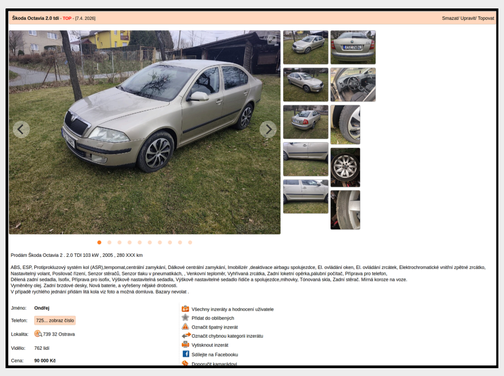


Fig. 1

Hyperinzerce.cz

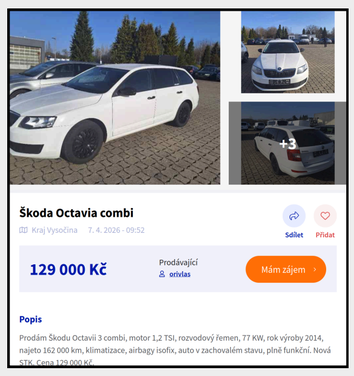


Fig. 2

Colly

Web Scraping Performance: Colly vs. BeautifulSoup

	Brand	Listings count	Average time
Go (Colly)	Mazda	2126	3.10 s
	Audi	7702	19.02 s
Python (BeautifulSoup)	Mazda	2126	13.24 s
	Audi	7702	50.19 s

Table 1

2. Information extraction

LLM extraction

```
Prodám Škoda Octavia 2. 2.0 TDI 103 kw , 2005 , 280 XXX km
```

→

```
{
  “brand”: “Škoda”,
  “model”: “Octavia”,
  “displacement”: 2.0,
  “power”: 103,
  “fuel_type”: “diesel”,
  “mileage”: “280000”
}
```

Fig. 3

Normalization

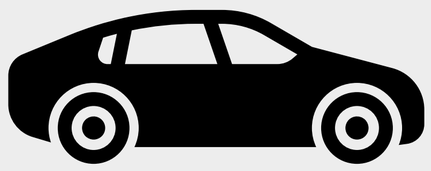
- 140 hp → 103 kw
- 1397 ccm → 1.4 l
- Oktavie RS → Octavia
- VW → Volkswagen
- TDI → diesel
- 280 xxx km → 280000

Qwen3

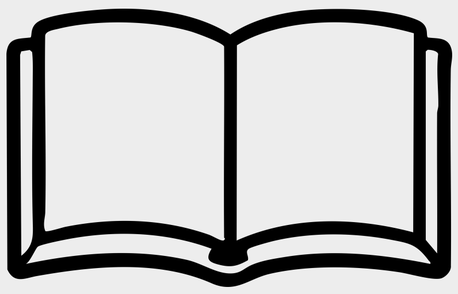
Fig. 4

3. Data extending

Extracted basic information



Cars catalog



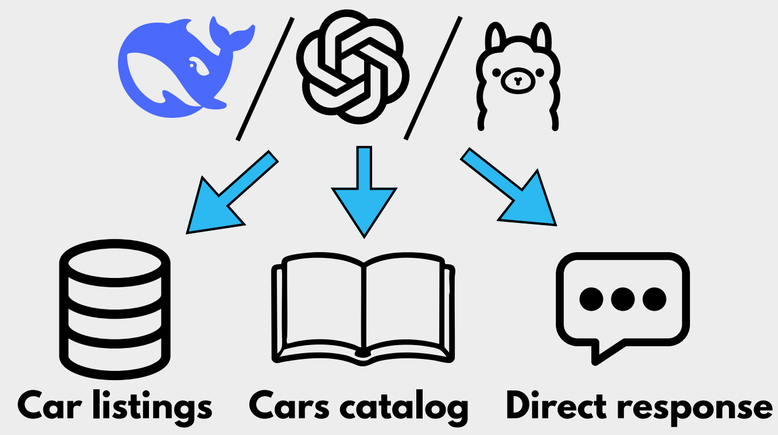
Enriched vehicle profile

- Škoda Octavia
- 103 kw
- FWD
- 5.5 l / 100 km
- 9.7 s
- 2004-2008
- 1425 kg

Fig. 5

4. AI assistant

Retrieval-Augmented Generation (RAG)



Strategies:

- SQL
- Semantic
- Hybrid
- Catalog
- Aggregate
- Clarify
- Conversational

BAAI/bge-m3

LangGraph

LangChain

Fig. 6

5. Web application

Chat interface

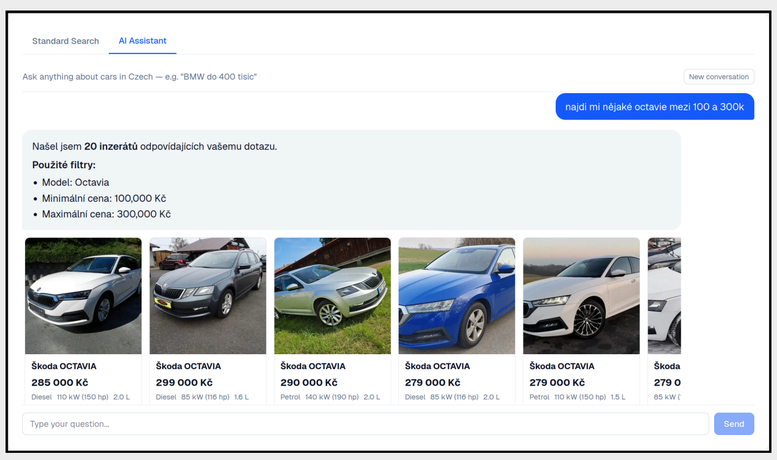


Fig. 7

Information Extraction Accuracy by LLM and Prompt Type

	Zero-shot	Persona	Few-shot Persona	CoT Persona	Few-shot CoT Persona
Qwen2.5:3B	81.89 %	85.18 %	86.64 %	79.79 %	80.07 %
Qwen2.5:7B	88.39 %	87.57 %	88.71 %	88.54 %	88.96 %
Qwen2.5:14B	91.29 %	92.68 %	93.04 %	92.39 %	92.11 %
Qwen2.5:32B	92.64 %	93.54 %	92.46 %	93.43 %	93.39 %
Qwen3:30B	90.18 %	89.93 %	92.89 %	93.25 %	93.79 %

Table 2