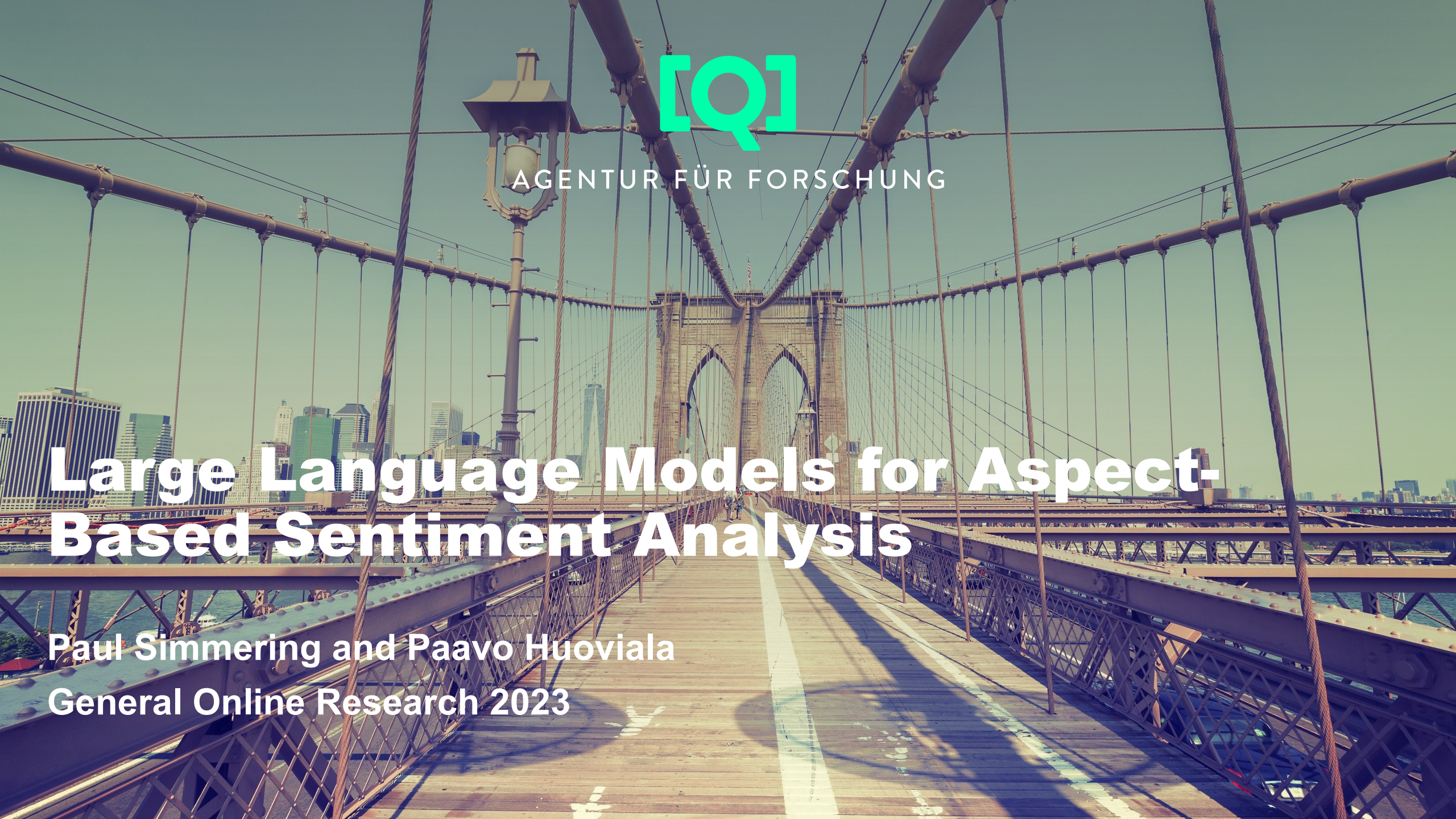


Large Language Models for Aspect-Based Sentiment Analysis

Paul Simmering and Paavo Huoviala

General Online Research 2023

Aspect-based sentiment analysis unlocks customer insights

"The gaming laptop's **GPU** is fantastic and runs all my games smoothly. Lots of **hard drive** space. But it took 2 months for **delivery** ☹"

Did large language models solve NLP?

Internet text

Books

Papers

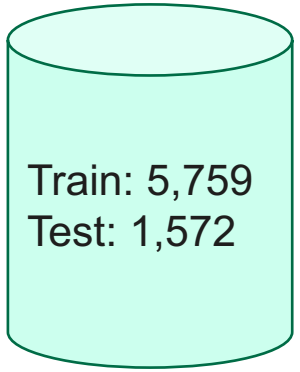
Code

The sky is _____

1. **Best answer**
2. **Second best**
3. **…**

More than 3B pages of input → Text completion → Reinforcement learning from human feedback → General task solver

Experiment setup



SemEval 2014
Restaurant +
Laptop review
Sentences
Pontiki et al., 2014

Component	Role	Content
Instruction (7 variants) 0 to 10 examples	System	You are an expert market researcher… The hard drive is fast {“term”: “hard drive”, “polarity”: “positive”} …
Input text	User	The sound drivers are buggy

Parameters: temperature = 0, others default

JSON Schema
Function definition

OpenAI

ChatCompletion API

Models:
- gpt-4-0613
- gpt-3.5-turbo-0613

```
{
  “aspects”: [
   {“term”: “sound drivers,
    “polarity”: “negative”}
   ]
}
```

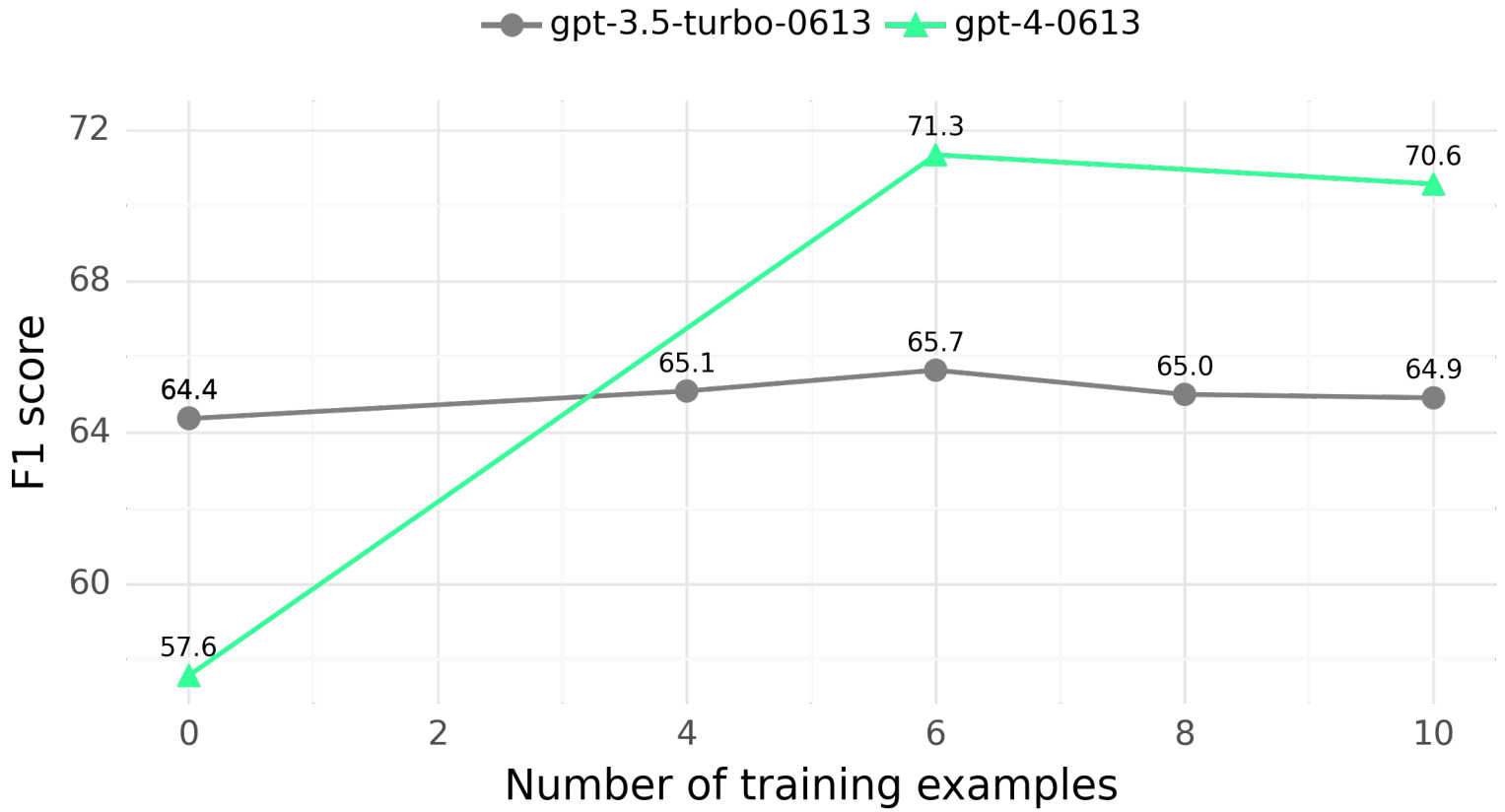
Data → API Requests → Models

Prompt comparison with gpt-3.5

Prompt	Description	Tokens	F1, zero examples	F1, 10 examples
Guidelines summary	Summary of the guidelines created by GPT-4	178	**64.39**	**64.92**
Annotation guidelines	Official guide for SemEval2014 Task 4	2021	63.77	64.88
Roleplay	Pretend to be a specialized machine learning model	84	62.09	64.64
Reference	Name-drop SemEval2014 Task 4	39	62.36	63.23
InstructABSA with examples	InstructABSA prompt + 6 examples from paper	249	61.54	62.18
InstructABSA	InstructABSA (SOTA) prompt	18	52.94	61.32
Separate tasks	2 steps: Term extraction, polarity classification	150	61.68	61.27

14 runs with gpt-3.5-turbo-0613 on 1572 test examples, pooled restaurants + laptops

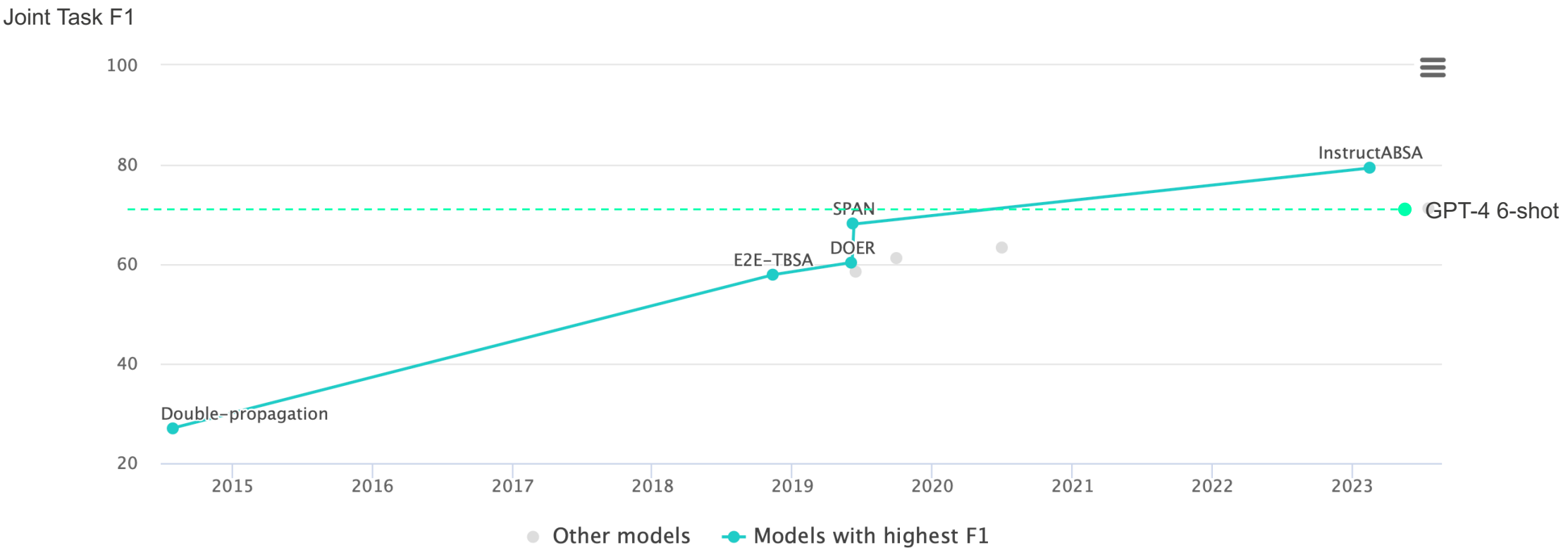
GPT-4 has strong in-context learning



- Hard examples not helpful
- InstructABSA F1: 79.3

Run with gpt-4-turbo-0613 and gpt-3.5-turbo-0613 on 1572 test examples, pooled restaurants + laptops, guidelines summary prompt

GPT-4 with few-shot instructions is competitive



https://paperswithcode.com/sota/aspect-based-sentiment-analysis-on-semeval-6

Benchmark errors are rarely outright mistakes

Error type	%	Example prediction	Gold answer
Predicted aspect not in gold aspects	50%	*It's **fast**, **quiet**, incredibly **small** and **affordable** […]*	No aspects
Aspect boundaries	24%	*The **Mini's body** hasn't changed since late 2010- and for a good reason*	**body**
Wrong polarity	24%	*I called Toshiba […] they were having issues with the **mother boards***	**mother boards**
Made up aspect words	2%	*It is EXTREMELY fast and never lags (pred: **speed**, **performance**)*	No aspects

Total: 748 errors in 1572 examples, each example can have multiple aspects, GPT-4, 10 examples

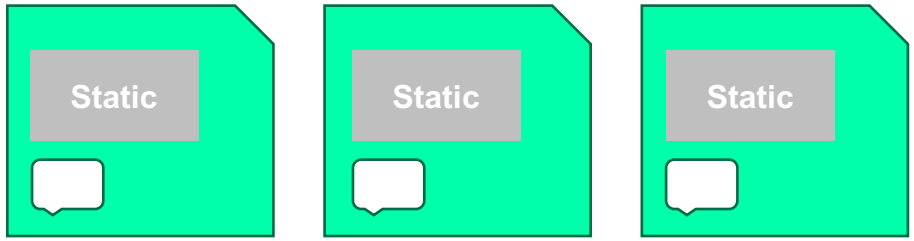
LLM Economics

Token usage

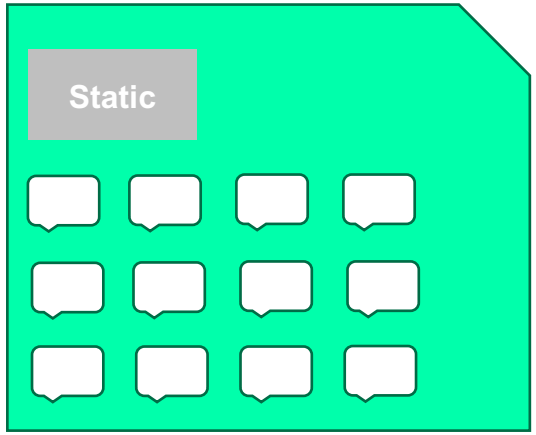
- System message + function definition: 818
- Average input sentence: 17
- Average output JSON: 26

Model	Parameters	Cost / 1000 sentences
GPT-4	$1.7x10^{12}$	$26.61
GPT-3.5	$1.8x10^{11}$	$1.60
InstructABSA (T5)	$2x10^{8}$	< $0.02

1 input text per request



Binpacked requests



More than 3x efficiency gain for small texts

LLM pipelines: Quick to set up, expensive to run

Step	Specialized model	LLM
Label	Label thousands of high-quality examples	Not required
Train	Hyperparameter tuning	Prompt tuning, few-shot examples, optional fine tuning
Inference	Efficient	Expensive, rate limited
Main expense	**Work time**	**Inference cost**

Python package texttunnel

- Simultaneous requests
- Binpacking
- Caching
- Cost estimation
- Open source, MIT license

Package: https://github.com/qagentur/texttunnel

Company: https://teamq.de

Twitter: https://twitter.com/q_insightagency

Paper & code forthcoming