KNOWLEDGE EXTRACTION WITH OPEN-SOURCE LLMS | KNEON

FINAL PRESENTATION

Francisco Monteiro | 03771311
Ismet Buyar | 03701344
Phillip Hauck | 03662213

AGENDA

01 BUSINESS UNDERSTANDING

02 DATA UNDERSTANDING & PREPARATION

03 MODELLING

04 EVALUATION

05 DEPLOYMENT

06 CONCLUSION

AGENDA

01 BUSINESS UNDERSTANDING

02 DATA UNDERSTANDING & PREPARATION

03 MODELLING

04 EVALUATION

05 DEPLOYMENT

06 CONCLUSION

Phenomenon

E.ON is one of Europe's largest operators of energy networks, infrastructure, and customer solutions

Large company

74K
Employees

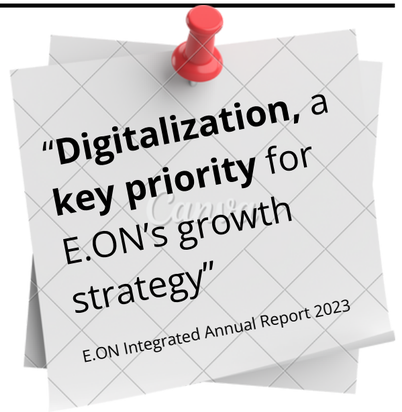
47M
Customers

+15
Countries

Digitalized company

40%
Customers served via digital platforms

14M
Smart energy meter instalations



Source: [2] E.ON, "Integrated Annual Report 2023," 2023. [Online]. Available: https://www.eon.com/en/about-us/annual-report-2023.html

Phenomenon

Its large and digitalized characteristics result in a overwhelming amount of digital textual data from both customer interactions and internal sources

Large
+
Digitalized

=

Overwhelming influx / generation of:

Customer feedback

Internal documents

=

How to deal with this challenge

Phenomenon

Companies are using LLM commercial solutions for Knowledge Extraction tasks

Provider	OpenAI Enterprise	Microsoft Azure	Amazon Bedrock	Google Vertex
LLM Models	GPT-4o, GPT-4 Turbo, GPT-4, GPT-3.5 Turbo Instruct, GPT-3.5 Turbo	GPT-4 Turbo, GPT-4, GPT-3.5 Turbo Instruct, GPT-3.5 Turbo, Llama 3 (70B), Llama 3 (8B), Llama 2 Chat (70B), Llama 2 Chat (13B), Llama 2 Chat (7B), Mistral Large, Mistral Small, Command-R+, Command-R	Llama 3 (70B), Llama 3 (8B), Llama 2 Chat (70B), Llama 2 Chat (13B), Mistral Large, Mixtral 8x7B, Mistral 7B, Claude 3 Opus, Claude 3 Sonnet, Claude 3 Haiku, Claude Instant, Claude 2.1, Command Light, Command	Gemini 1.5 Pro, Gemini 1.5 Flash, Gemini 1.0 Pro, Claude 3 Sonnet, Claude 3 Haiku
Companies using	moderna, Klarna., Harvey.	e.on, IVECO GROUP, pwc	adidas, Alida., lonely planet, CLARIANT	Audiomob, SILVR, Mercedes

Other Providers

Anthropic

Groq

Together.ai

Mistral

Use cases:

- Smart internal chatbot (for engineers and other employees) for Q&A / extract information from internal data and documents (Iveco Group, Adidas, Clariant, Pwc)
- Sentiment Analysis on gathered customer feedback (Alida)
- Enhancing products/services by leveraging LLM's (Lonelyplanet)

Source: [6] Artificial Analysis, "Artificial Intelligence and Analysis," [Online]. Available: https://artificialanalysis.ai/. [Accessed: Jul. 9, 2024].
[8] OpenAI, "Stories," [Online]. Available: https://openai.com/news/stories/. [Accessed: Jul. 9, 2024]; [9] Microsoft, "AI Customer Stories," [Online]. Available: https://www.microsoft.com/en-us/ai/ai-customer-stories. [Accessed: Jul. 9, 2024].
[10] Amazon Web Services, "Bedrock Testimonials," [Online]. Available: https://aws.amazon.com/bedrock/testimonials/. [Accessed: Jul. 9, 2024].

Problem

E.ON is currently facing critical challenges regarding commercial NLP solution

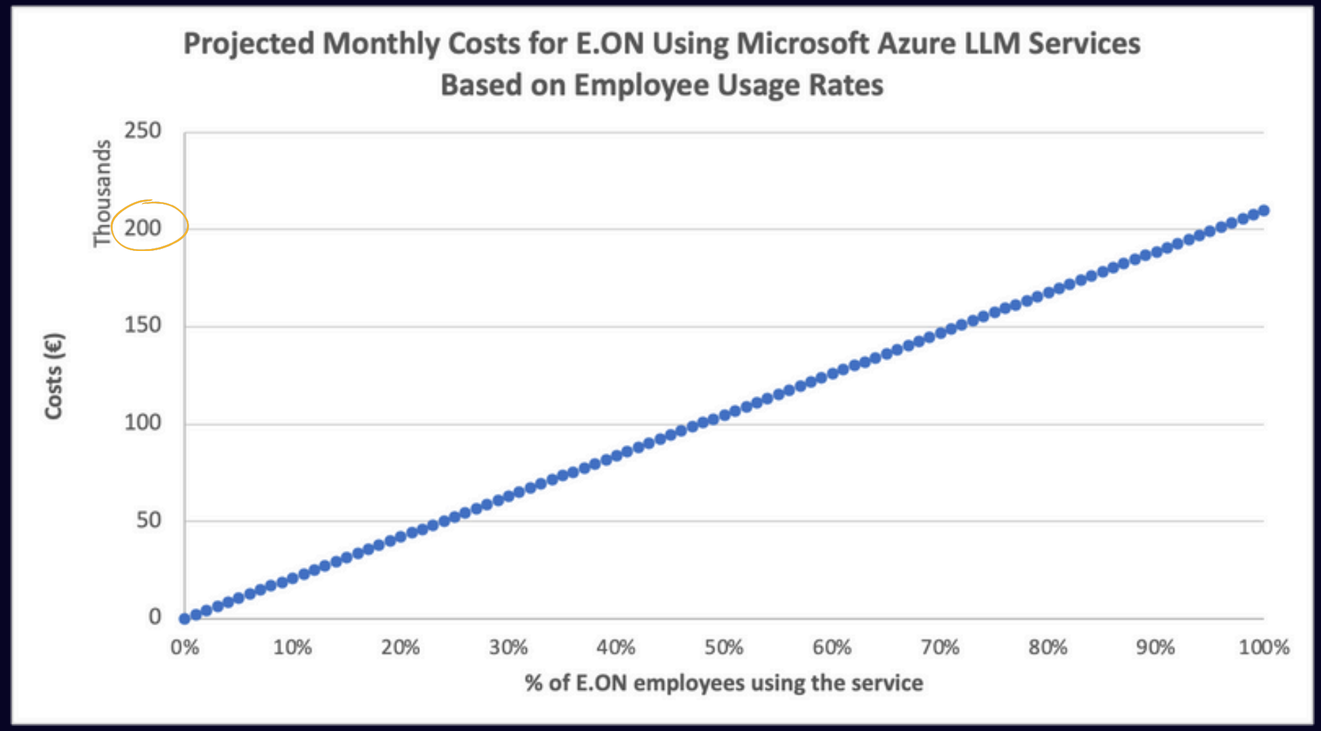
E.ON requirements	Commercial Solutions
Data privacy & security	✓
Ease of Integration	✓
Cost-effectiveness	✗
Performance Models quality	✓
Performance User experience	✗
Customization Fine-tuning, Multimodality	✗

Source: [7] HatchWorks, "Open Source vs Closed LLMs: A Comprehensive Guide," [Online]. Available: https://hatchworks.com/blog/gen-ai/open-source-vs-closed-llms-guide/. [Accessed: Jul. 9, 2024].

Problem

E.ON is currently facing critical challenges regarding commercial NLP solution

E.ON requirements	Commercial Solutions
Data privacy & security	✓
Ease of Integration	✓
Cost-effectiveness	✗
Performance Models quality	✓
Performance User experience	✗
Customization Fine-tuning, Multimodality	✗



Assumptions:

Avg price per million tokens (blend of input+output): 12.88€
N° tokens generated per worker per month: 10 000
Working days per month: 22
Total number of workers: 74 000

Source: [7] HatchWorks, "Open Source vs Closed LLMs: A Comprehensive Guide," [Online]. Available: https://hatchworks.com/blog/gen-ai/open-source-vs-closed-llms-guide/. [Accessed: Jul. 9, 2024].
[6] Artificial Analysis, "Artificial Intelligence and Analysis," [Online]. Available: https://artificialanalysis.ai/. [Accessed: Jul. 9, 2024].

Opportunity

Leveraging open source LLM’s could be an alternative to commercial offerings

E.ON requirements	Commercial Solutions	Open-source solutions	Required resources
Data privacy & security	✓	✓!	Significant technical expertise to implement tailored security protocols
Ease of Integration	✓	✓!	Significant technical expertise on how to integrate with existing systems
Cost-effectiveness	✗	✓!	Significant technical expertise on customization, scaling and maintenance to maximize model's utility and security.
Performance Models quality	✓	✓!	Evaluate performance metrics accross different models and select the best ones
Performance User experience	✗	✓!	
Customization Fine-tuning, Multimodality	✗	✓!	Significant technical expertise to modify and maintain the models.

Source: [7] HatchWorks, "Open Source vs Closed LLMs: A Comprehensive Guide," [Online]. Available: https://hatchworks.com/blog/gen-ai/open-source-vs-closed-llms-guide/. [Accessed: Jul. 9, 2024].

Opportunity

Leveraging open source LLM’s could be an alternative to commercial offerings

E.ON requirements	Commercial Solutions	Open-source solutions	Required resources
Data privacy & security	✓	✓!	Significant technical expertise to implement tailored security protocols
Ease of Integration	✓	✓!	Significant technical expertise on how to integrate with existing systems
Cost-effectiveness	✗	✓!	Significant technical expertise on customization, scaling and maintenance to maximize model's utility and security.
Performance Models quality	✓	✓!	Evaluate performance metrics accross different models and select the best ones
Performance User experience	✗	✓!	
Customization Fine-tuning, Multimodality	✗	✓!	Significant technical expertise to modify and maintain the models.

Source: [7] HatchWorks, "Open Source vs Closed LLMs: A Comprehensive Guide," [Online]. Available: https://hatchworks.com/blog/gen-ai/open-source-vs-closed-llms-guide/. [Accessed: Jul. 9, 2024].

Focus of this project

Use case

Question answering (knowledge retrieval) on unstrucutred textual sources

Project main task

Optimisation of available open-source LLMs + evaluation of results with appropriate metrics

Datasets

SQuAD1.1, SQuAD2.0, GermanQuAD

Models

Bert-base-cased, Roberta-large, Albert-v2-large, Distillbert-base-uncased, Roberta-base

AGENDA

01 BUSINESS UNDERSTANDING

02 DATA UNDERSTANDING & PREPARATION

03 MODELLING

04 EVALUATION

05 DEPLOYMENT

06 CONCLUSION

Data Understanding

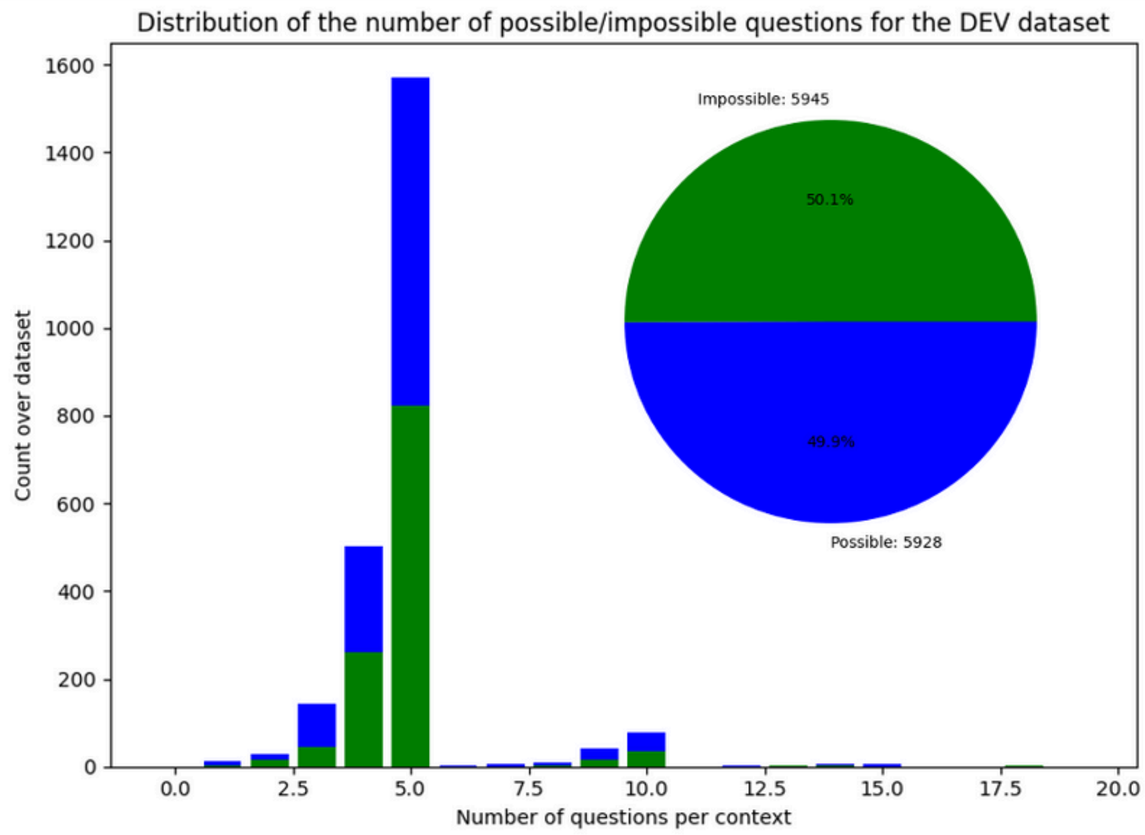
The Stanford Question Answering Dataset (SQuAD):

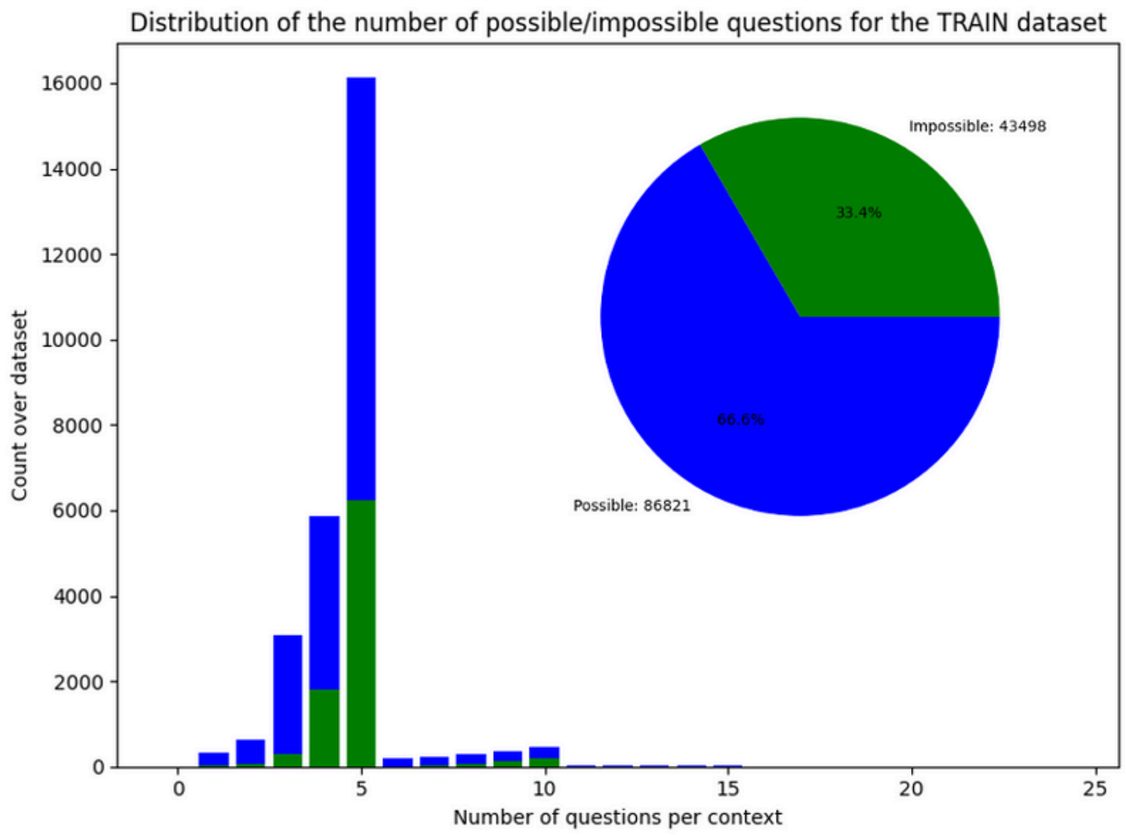
- Established benchmark for machine reading comprehension tasks
- First introduced 2016
- 100.000 question-answer pairs where the answer is in one of 20.000 segments from a Wikipedia passage (context)
- Primarily focused on factual, short-answer retrieval

Data Understanding

SQuAD 2.0:

- Introduces impossible questions
- Adds more than 50.000 questions to the existing ones

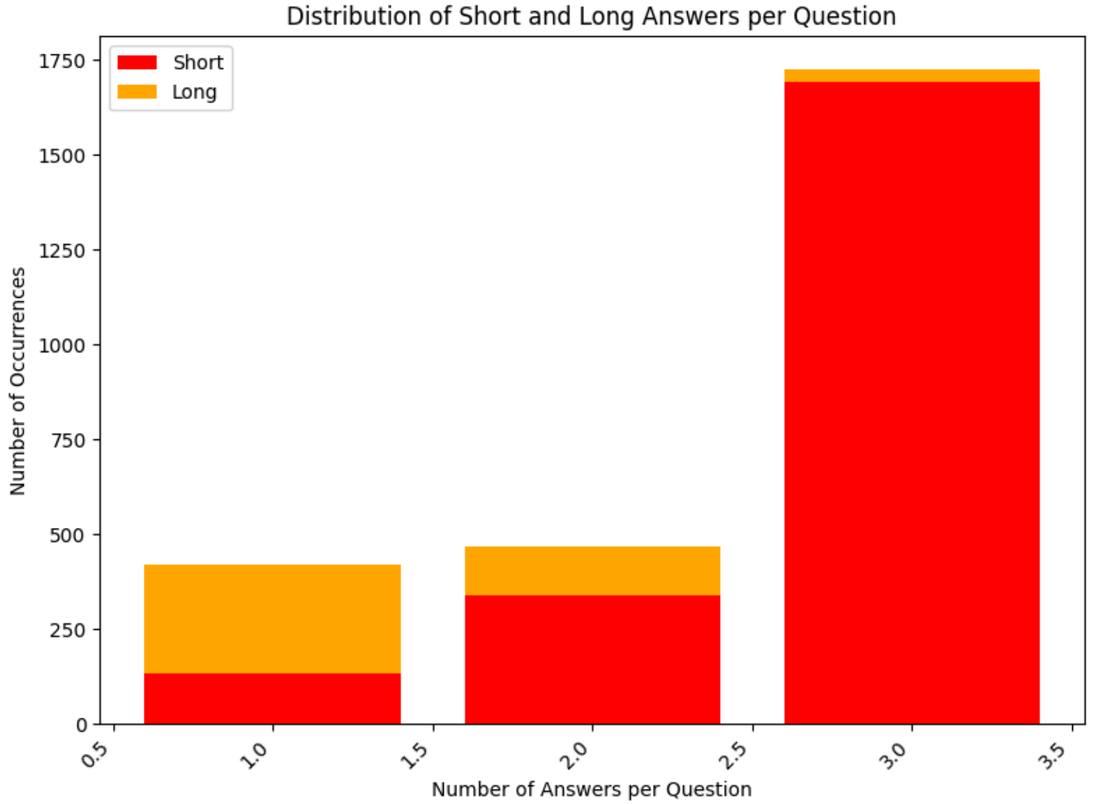




Data Understanding

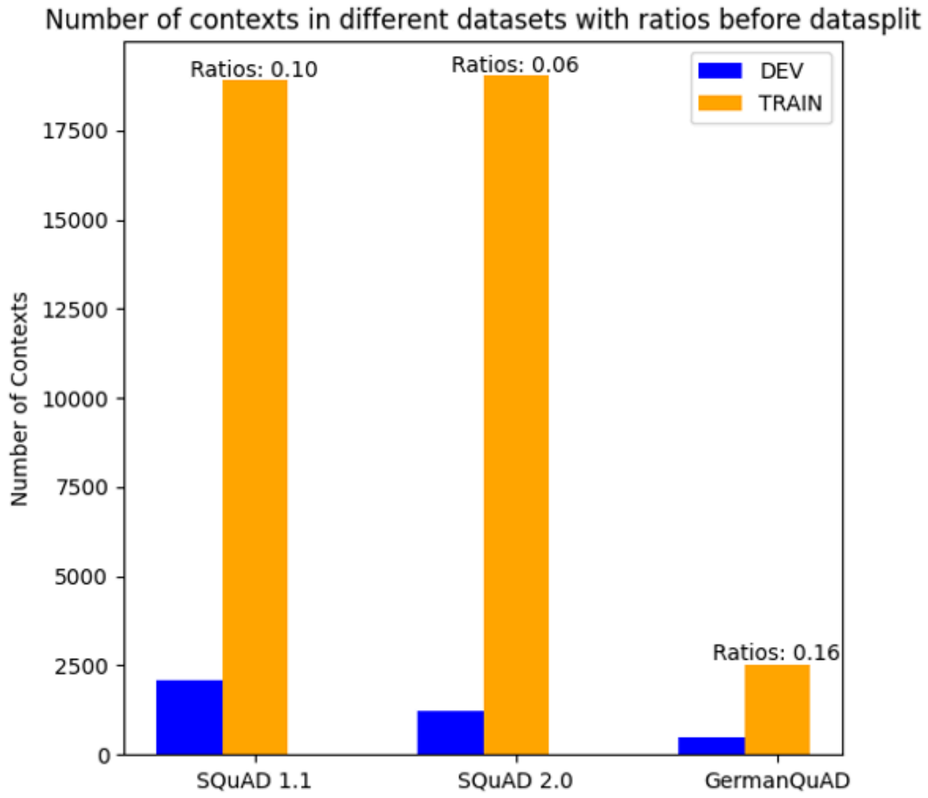
GermanQuAD:

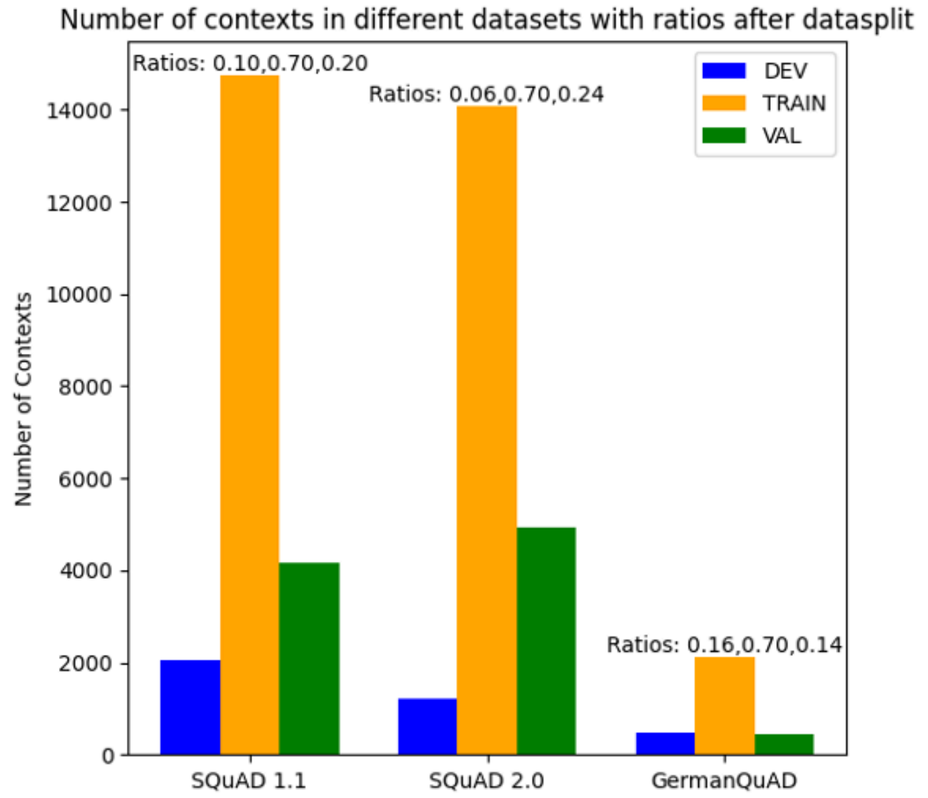
- German dataset inspired by SQuAD
- Significantly smaller number of contexts, but they are longer
- Incorporates "self-sufficient questions"
- Add distinction between long and short answers



Data Preparation

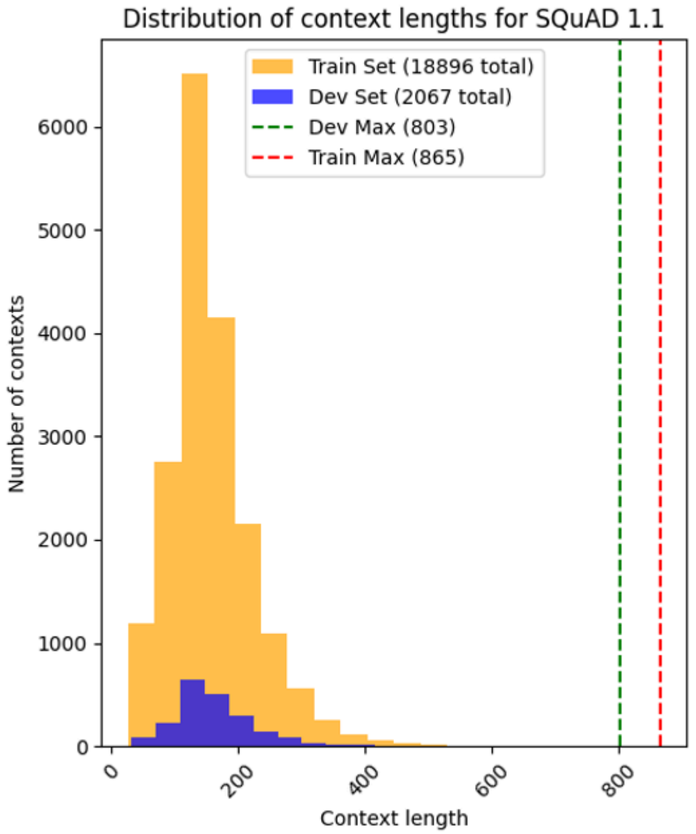
1. Cleaning of GermanQuad for data analysis
2. Datasplit (following ratios of possible/impossible questions for the SQuAD 2.0 dataset)

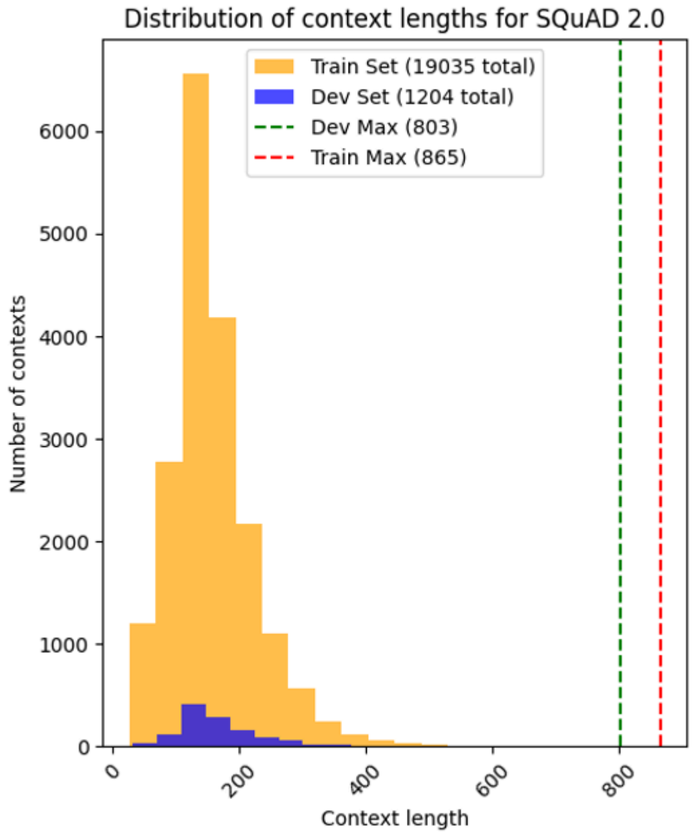


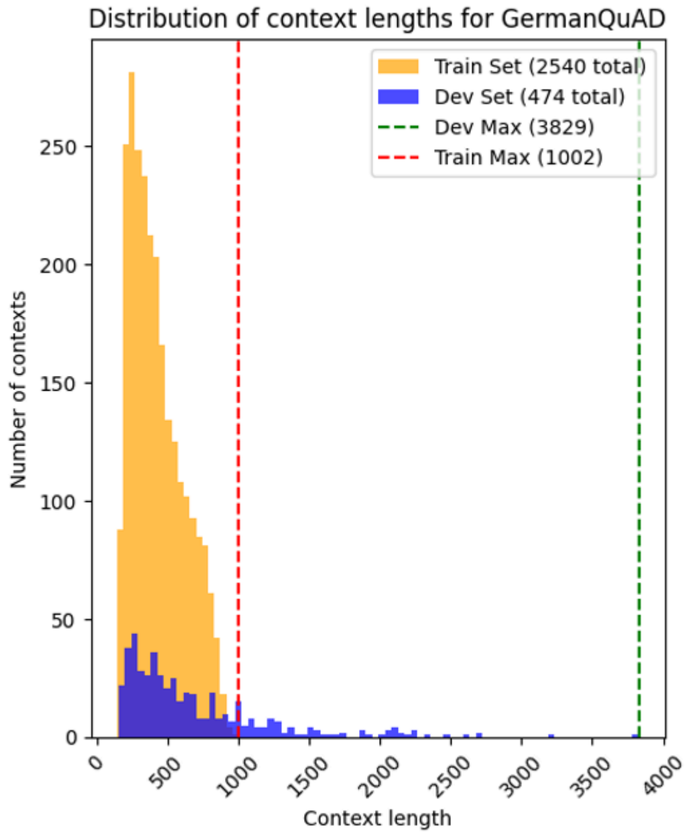


Data Preparation

3. Truncation:
   - Initially not required for all models
   - All datasets exceed bert-based model max token length







AGENDA

01 BUSINESS UNDERSTANDING

02 DATA UNDERSTANDING & PREPARATION

03 MODELLING

04 EVALUATION

05 DEPLOYMENT

06 CONCLUSION

Experiments

Evaluation with Exact Match (EM), F1 and BLEU Score

Raw Performance of BERT Family Models on SQUAD1.1

- 5% Random Sampling

Hyperparameter Optimization of BERT-base-cased on SQUAD1.1

- 5% Random Sampling

Raw Performance of BERT Family Models on SQUAD2.0

- 5% Stratified Sampling

Hyperparameter Optimization of BERT-base-cased on SQUAD2.0

- 5% Stratified Sampling

Raw performance of German-finetuned Models on GermanQUAD

- 10% Stratified Sampling

BERT Family Models for SQUAD1.1 and SQUAD2.0

Subtitle

BERT-base-cased (110M parameters)

DistilBERT-base-uncased (66M parameters)

- Distilled version of BERT

RoBERTa-base (125M parameters)

- Balanced performance and efficiency

RoBERTa-large (355M parameters)

ALBERT-v2-large (18M parameters)

- Compact and parameter-efficient
- Suitable for memory-constraint scenarios

German-finetuned Models for GermanQUAD

Subtitle

GElectra-base (110M parameters)

- German version of Electra
- Efficiently pre-trained with replaced token detection

GBERT-base (110M parameters)

- German BERT model

Hyperparameter Optimization Algorithms

Random Search

- Randomly samples hyperparameters
- Explores a wide range

Grid Search

- Exhaustively searches a predefined grid
- May miss better combinations outside the grid

Tree-structured Parzen Estimator (TPE)

- Sequential model-based optimization
- Adapts the search space dynamically

Hyperparameters

Baseline

"num_train_epochs": 3
"learning_rate": 3e-5
"weight_decay": 0.01
"per_device_train_batch_size": 8
"per_device_eval_batch_size": 8

Search Space (Optuna Backend, 3 Trials)

"num_train_epochs": [2, 3, 4, 5]
"learning_rate": 1e-5 - 5e.5
"weight_decay": 0.005 - 0.02
"per_device_train_batch_size": [8, 16, 32]
"per_device_eval_batch_size": [8, 16, 32]

AGENDA

01 BUSINESS UNDERSTANDING

02 DATA UNDERSTANDING & PREPARATION

03 MODELLING

04 EVALUATION

05 DEPLOYMENT

06 CONCLUSION

Raw Performance of BERT Family Models on SQUAD1.1

	EM	F1	BLUE
BERT-base-cased	0.62	0.73	0.03
DistilBERT-base-uncased	0.54	0.66	0.09
RoBERTa-base	0.74	0.84	0.15
RoBERTa-large	0.82	**0.90**	-
ALBERT-v2-large	0.78	0.87	-

Hyperparameter Optimization of BERT-base-cased on SQUAD1.1

	EM	F1	BLUE
Baseline	0.62	0.73	0.03
Best Random	0.60	0.71	0.03
Best Grid	0.62	0.73	0.03
Best TPE	0.62	0.73	0.03

Best TPE Hyperparameters

{'num_train_epochs': 2,
'learning_rate': 1.0153553712098332e-05,
'weight_decay': 0.019536412032851 32,
'per_device_train_batch_size': 8,
'per_device_eval_batch_size': 8}

Raw Performance of BERT Family Models on SQUAD2.0

	EM	F1	BLUE
BERT-base-cased	0.49	0.52	0.09
DistilBERT-base-uncased	0.46	0.47	0.04
RoBERTa-base	0.59	0.63	0.18
RoBERTa-large	0.78	0.82	0.18
ALBERT-v2-large	0.73	0.78	0.19

Hyperparameter Optimization of BERT-base-cased on SQUAD2.0

	EM	F1	BLUE
Baseline	0.49	0.52	0.09
Best Random	0.51	0.54	0.09
Best Grid	0.51	0.54	0.15
Best TPE	0.54	0.52	0.12

Best Grid Hyperparameters

{'num_train_epochs': 4,
'learning_rate': 5e-05,
'weight_decay': 0.02,
'per_device_train_batch_size': 8,
'per_device_eval_batch_size': 16}

Raw performance of German-finetuned Models on GermanQUAD

	EM	F1	BLUE
GElectra-base	0.19	0.40	0.19
GBERT-base	0.08	0.21	0.09

AGENDA

01 BUSINESS UNDERSTANDING

02 DATA UNDERSTANDING & PREPARATION

03 MODELLING

04 EVALUATION

05 DEPLOYMENT

06 CONCLUSION

Deployment Infrastructure (not enterprise-ready)

- Streamlit (Frontend)
- Django (Backend)
- Docker Compose (Orchestration)
- Private HuggingFace Model Registry (Model Hub)

AGENDA

01 BUSINESS UNDERSTANDING

02 DATA UNDERSTANDING & PREPARATION

03 MODELLING

04 EVALUATION

05 DEPLOYMENT

06 CONCLUSION

Conclusions

Future Research

- Hyperparameter Optimization on other BERT Models
- Parameter-efficient Finetuning (PEFT)

Limitations

- Not all Models have Question Answering Head in HuggingFace
- Compute, Memory and Time Constraints

References

[1] K. Chavez, “Quantized Mistral 7B vs TinyLlama for Resource-Constrained Systems,” Towards Data Science, May 29, 2023. [Online]. Available: https://towardsdatascience.com/quantized-mistral-7b-vs-tinyllama-for-resource-constrained-systems-a6ce4ab95b03

[2] E.ON, “Integrated Annual Report 2023,” 2023. [Online]. Available: https://www.eon.com/en/about-us/annual-report-2023.html

[3] L. Wang, M. Feng, B. Zhou, B. Xiang, and M. Sridhar, “Efficient hyper-parameter optimization for NLP applications,” in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 2015, pp. 2112–2117. DOI: 10.18653/v1/D15-1253.

[4] Y. Wijeratne and F. Marikar, “Better Question-Answering Models on a Budget,” arXiv, Apr. 2023. [Online]. Available: https://arxiv.org/abs/2304.12370

[5] Y. Zhang and Z. Xu, “BERT for Question Answering on SQuAD 2.0,” Stanford University. [Online]. Available: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/default/15848021.pdf

[6] Artificial Analysis, "Artificial Intelligence and Analysis," [Online]. Available: https://artificialanalysis.ai/. [Accessed: Jul. 9, 2024].

[7] HatchWorks, "Open Source vs Closed LLMs: A Comprehensive Guide," [Online]. Available: https://hatchworks.com/blog/gen-ai/open-source-vs-closed-llms-guide/. [Accessed: Jul. 9, 2024].

[8] OpenAI, "Stories," [Online]. Available: https://openai.com/news/stories/. [Accessed: Jul. 9, 2024].

[9] Microsoft, "AI Customer Stories," [Online]. Available: https://www.microsoft.com/en-us/ai/ai-customer-stories. [Accessed: Jul. 9, 2024].

[10] Amazon Web Services, "Bedrock Testimonials," [Online]. Available: https://aws.amazon.com/bedrock/testimonials/. [Accessed: Jul. 9, 2024].