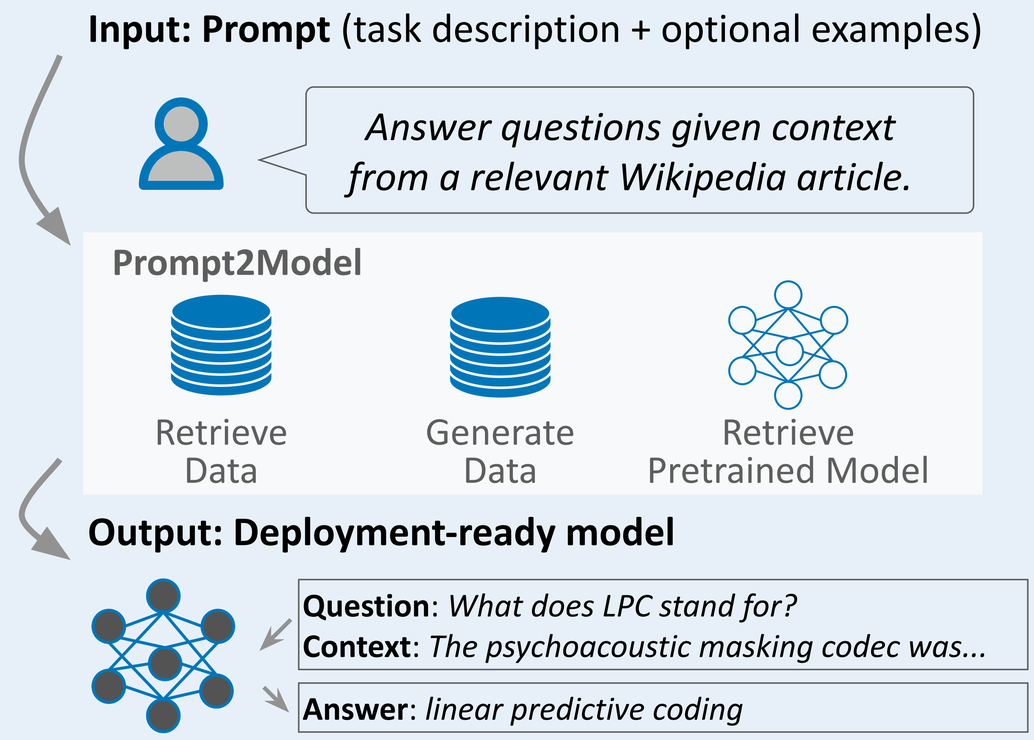
LLMs are expensive and have privacy concerns.

Prompt2Model creates **small, deployable expert models from prompts**, matching LLM performance in many cases.



Prompt2Model: Generating Deployable Models from Natural Language Instructions

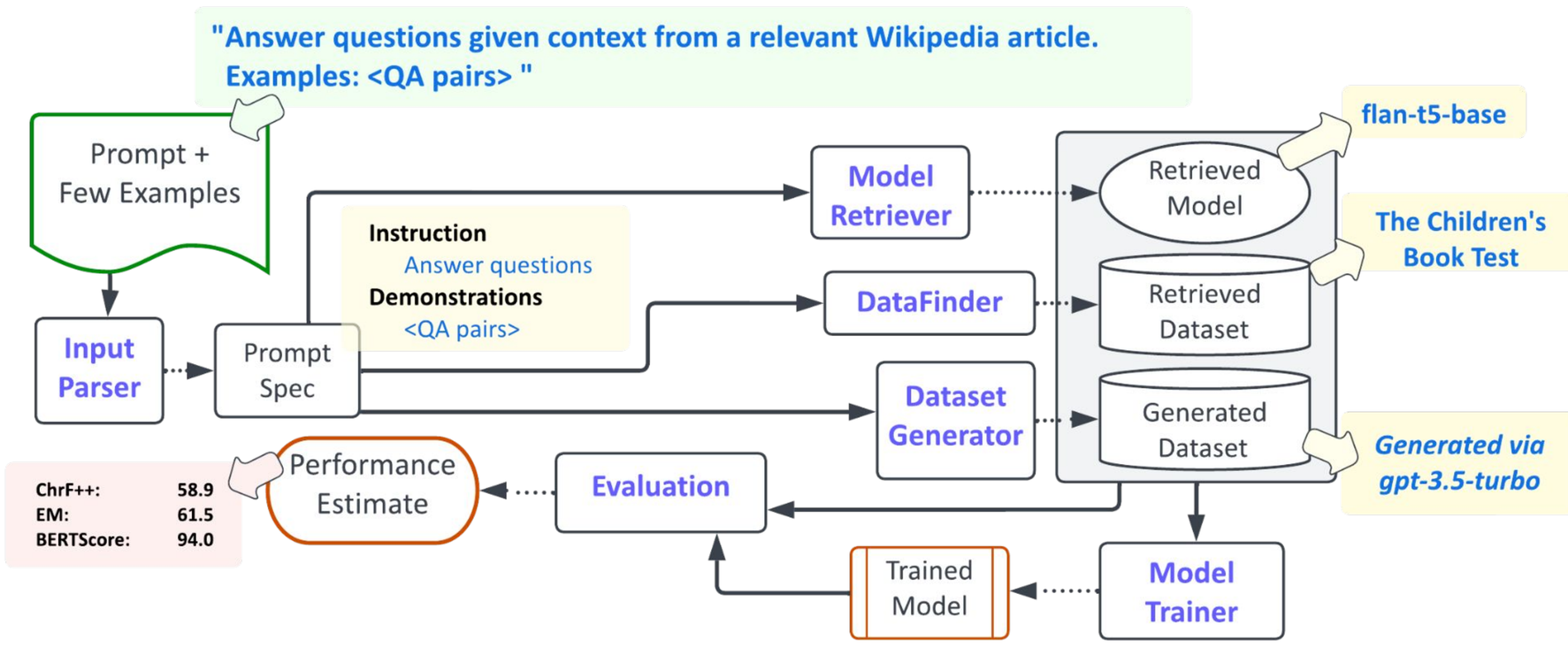
Background:

LLMs enable one-click NLP system design.

But they come with some practical downsides:

1. LLMs are expensive to serve (often must use an API)
2. They are slow
3. APIs are not private
4. They're hard to evaluate or fix

Our solution:



Results:

Small models from Prompt2Model can outperform large LMs.

Method	SQuAD (EM)	MCoNaLa (ChrF++)	Temporal (ChrF++)
Prompt2Model	61.5	13.1	55.2
w/o Model Ret.	61.5	15.8	55.2
w/o Data Ret.	50.2	16.6	N/A
`gpt-3.5-turbo`	42.1	37.3	30.7

For SQuAD, both retrieved and generated datasets provide value.

Method	#Train	Performance	Anno. Cost
Retrieval only	3,000	56.79	≈ $ 0
Generation only	3,000	44.20	≈ $ 5
Retrieval+generation	6,000	61.46	≈ $ 5
Custom annotation	3,000	61.64	≈ $ 540

Why should you care?

1. Deploying personal NLP models without APIs.
2. End-to-end AutoML with automatic data collection.
3. Fostering research in model distillation and dense retrieval.

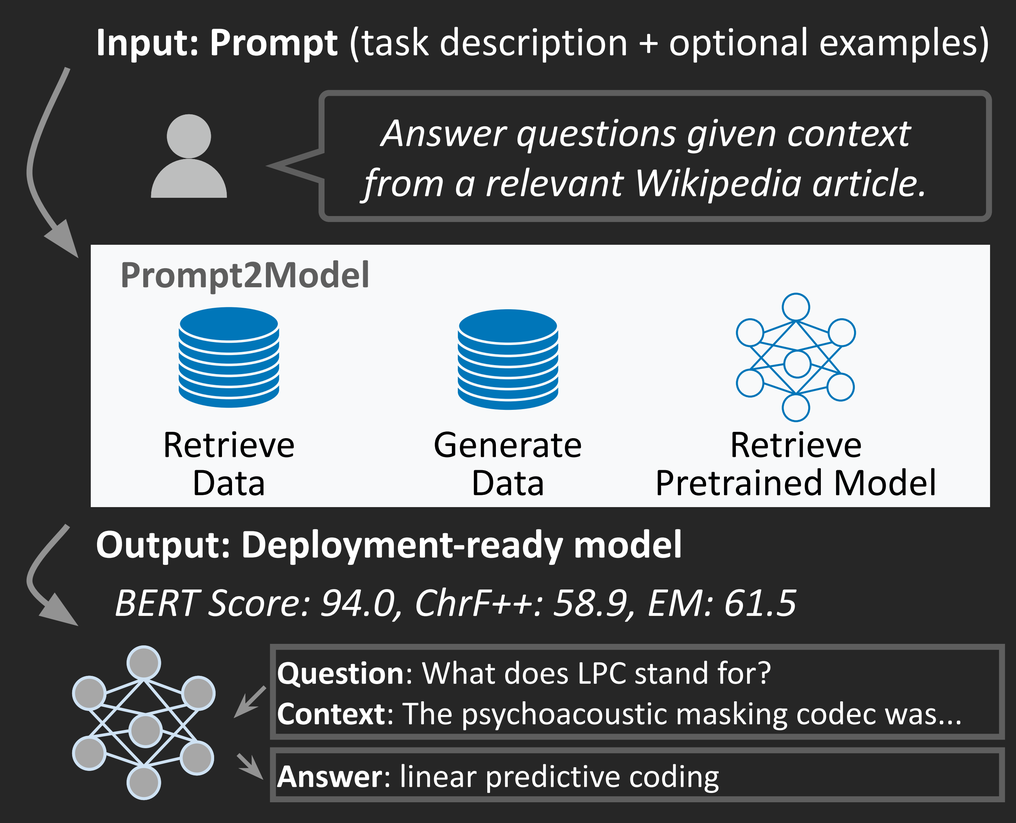
```
pip install prompt2model
neulab/prompt2model
```


Vijay Viswanathan*, Chenyang Zhao*,
Amanda Bertsch, Tongshuang Wu, Graham Neubig

Carnegie Mellon University
Language Technologies Institute
清華大學 Tsinghua University
北京信息科学与技术国家研究中心

LLMs are expensive and have privacy concerns.

Prompt2Model creates **small, deployable expert models from prompts**, matching LLM performance in many cases.



Prompt2Model: Generating Deployable Models from Natural Language Instructions

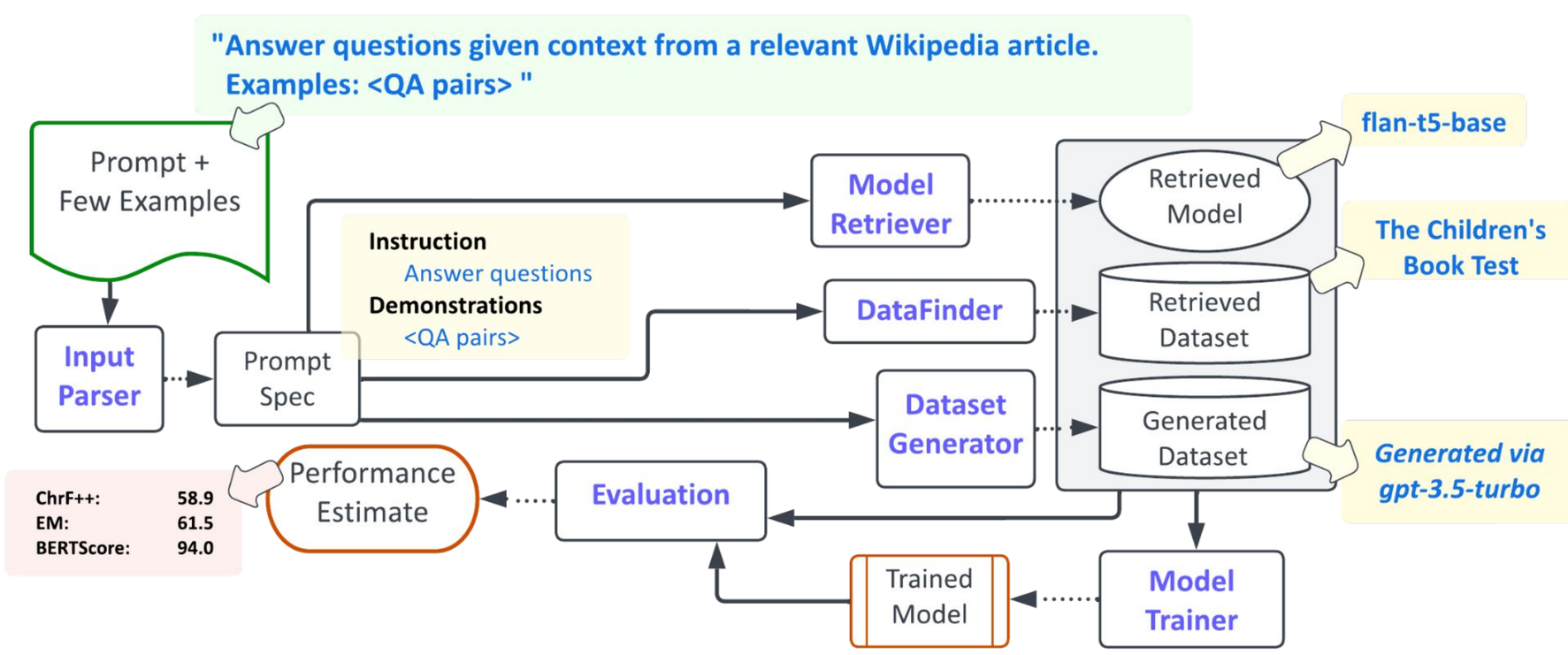
Background:

LLMs enable one-click NLP system design.

But they come with some practical downsides:

1. LLMs are expensive to serve (often must use an API)
2. They are slow
3. APIs are not private
4. They're hard to evaluate or fix

Our solution:



Results:

Small models from Prompt2Model can outperform large LMs.

Method	SQuAD (EM)	MCoNaLa (ChrF++)	Temporal (ChrF++)
Prompt2Model	61.5	13.1	55.2
w/o Model Ret.	61.5	15.8	55.2
w/o Data Ret.	50.2	16.6	N/A
`gpt-3.5-turbo`	42.1	37.3	30.7

For SQuAD, both retrieved and generated datasets provide value.

Method	#Train	Performance	Anno. Cost
Retrieval only	3,000	56.79	≈ $ 0
Generation only	3,000	44.20	≈ $ 5
Retrieval+generation	6,000	61.46	≈ $ 5
Custom annotation	3,000	61.64	≈ $ 540

Why should you care?

1. Deploying personal NLP models without APIs.
2. End-to-end AutoML with automatic data collection.
3. Fostering research in model distillation and dense retrieval.

```
pip install prompt2model
neulab/prompt2model
```


Vijay Viswanathan*, Chenyang Zhao*,
Amanda Bertsch, Tongshuang Wu, Graham Neubig

Carnegie Mellon University
Language Technologies Institute
Tsinghua University
北京信息科学与技术国家研究中心