Step1 Automotive Corpus Construction

QA corpus ➡ Documents Collection ➡ Multitask Annotated Corpus

Question: WEY是什么品牌的车?
(Which automobile brand is WEY?)
Answer: WEY是长城汽车旗下的高端品牌…
(WEY is a high-end brand under Great Wall Motors…)

Sentence 1: WEY是长城汽车旗下的高端品牌。
(WEY is a high-end brand under Great Wall Motors.)
Sentence 2: 作为早期产品，VV系列在动力匹配方面…
(As early products, the VV series faced challenges in powertrain matching…)

MLM: WEY是长[MASK]汽[MASK]旗下…
NSP: Sentence pair: (Sentence1, Sentence2)
　　　Label: continuous (1)
STS: Sentence pair: (Sentence1, Sentence2)
　　　Label: similar (1)
TE: 　Sentence pair: (Sentence1, Sentence2)
　　　Label: neural (1)

Step2 Encoder: A Pretraining-Multitask Finetuning Framework

Pretraining

Open-domain Corpus → BERT

Single-task Finetuning

MLM
NSP
STS
TE

Multitask Finetuning

MLM + NSP
MLM + STS
MLM + TE

Step3 Retriever based on Faiss

Vector Space

Vectors Loader → Index

Querying

Query vector

Query results (Ranking)

Index	20	201	55	...
Distance	12	18	27	...