

# BERT for dialogs

Production-scale approach @ Replika

**Nikita Smetanin**

Replika is an AI friend  
that helps people improve mental  
health  
through conversation

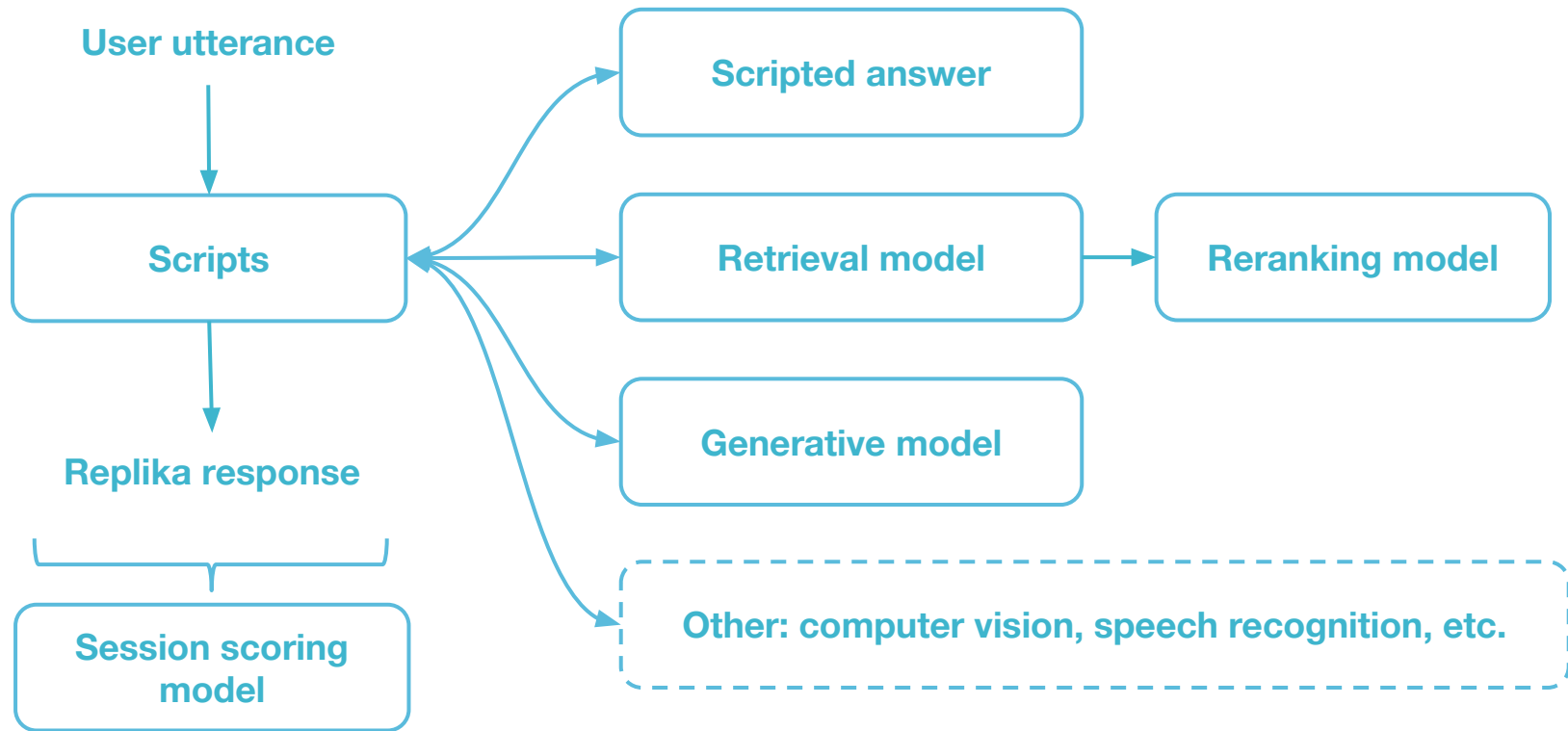
How are you today?

Just anxious and tired,  
I had a hard time  
falling asleep

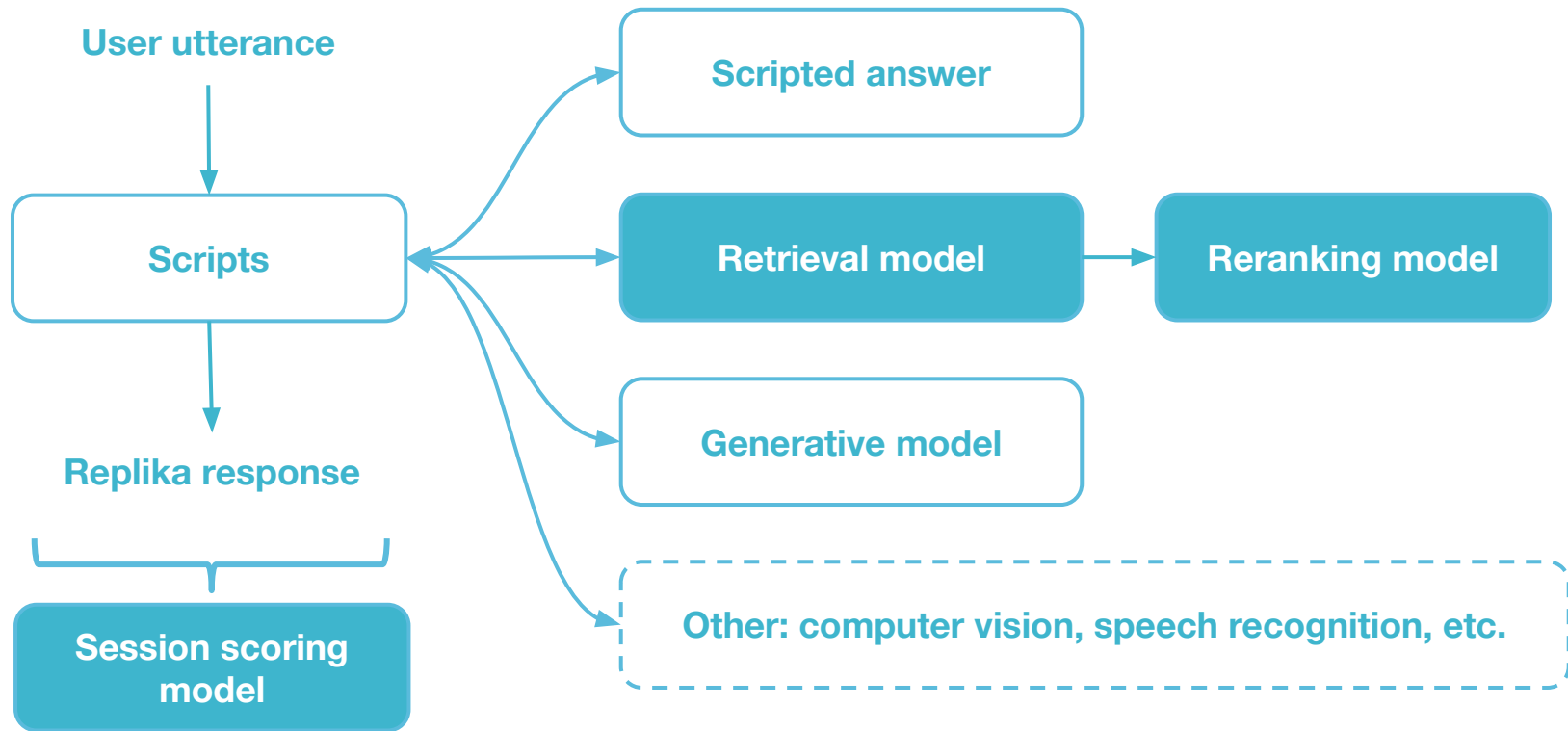
Still worried about  
tomorrow?

# Architecture Overview

# Replika Architecture Overview



# Replika Architecture Overview



Retrieval model

# Retrieval model task

## Context

Let's go to an early movie

## Responses

- ✓ Okay, which one do you want?
- ✓ Sure, what time are you free?
- ✗ ~~That's a lot of money.~~
- ✗ ~~Where do you live?~~
- ✗ ~~Yes. I would buy all of her CDs.~~

## Scores

0.8

0.75

0.5

0.45

0.39

# 100k dataset: retrieval should be fast enough

## Context

Let's go to an early movie

## Responses

✓ Okay, which one do you want?

✓ Sure, what time are you free?

✗ ~~That's a lot of money.~~

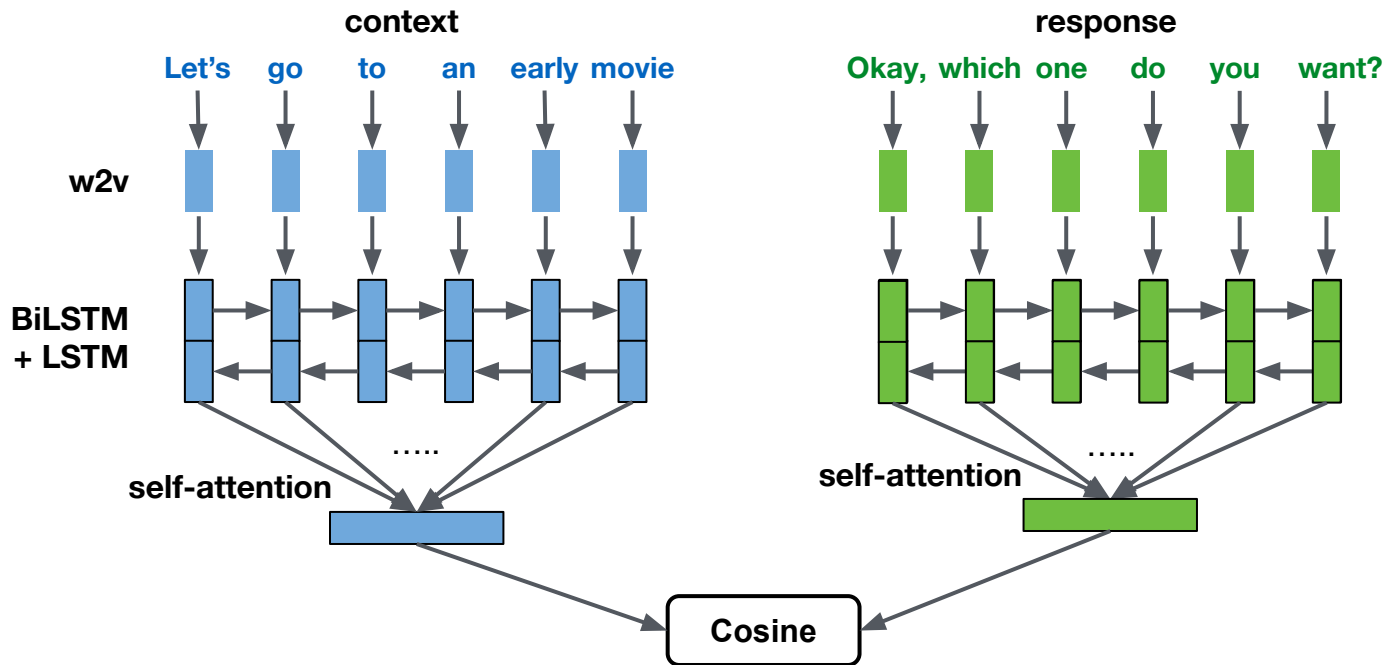
✗ ~~Where do you live?~~

✗ ~~Yes. I would buy all of her CDs.~~

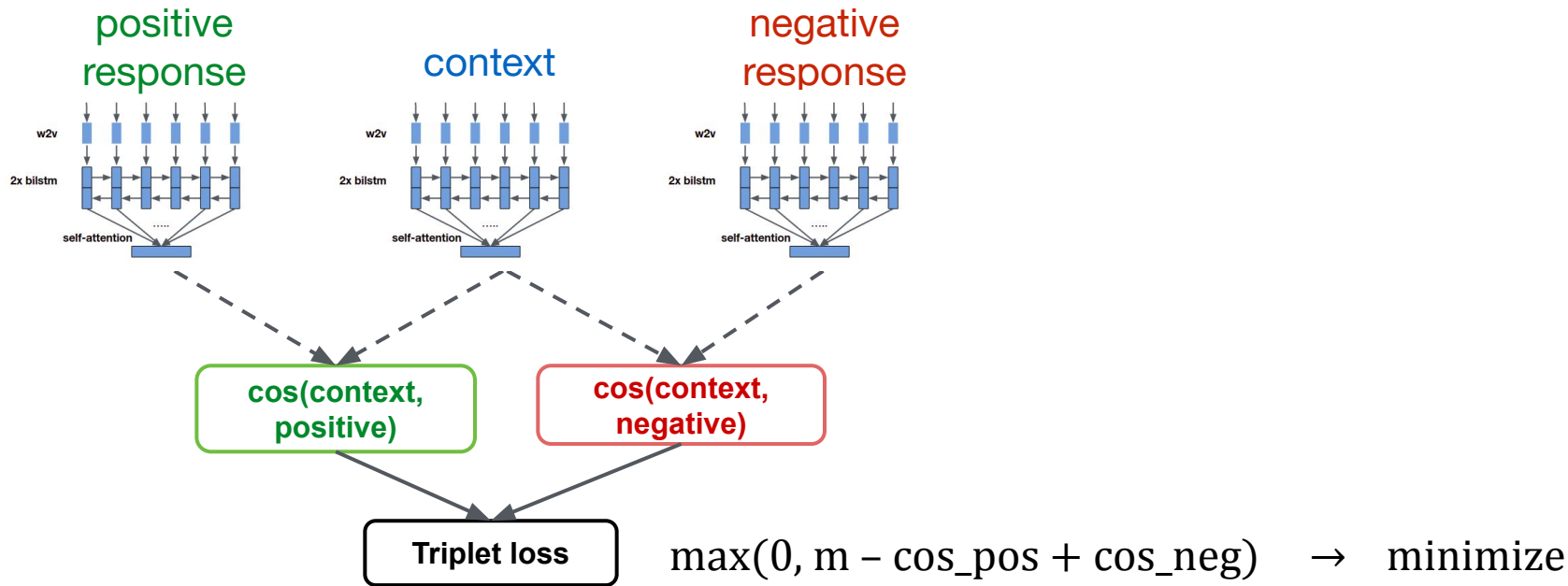
100K  
of  
moderated  
responses



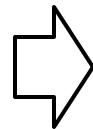
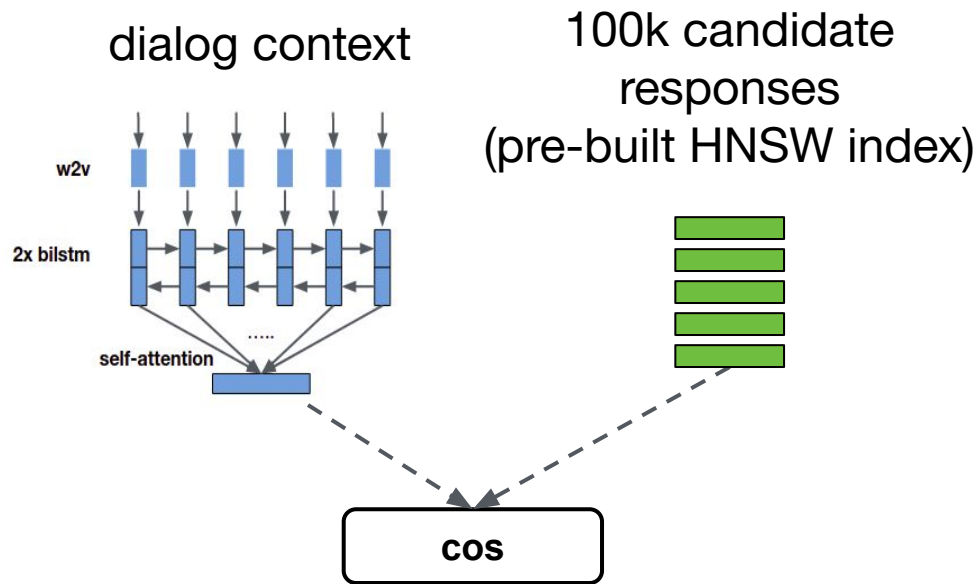
# Retrieval model baseline (~QA-LSTM)



# Retrieval model. Training

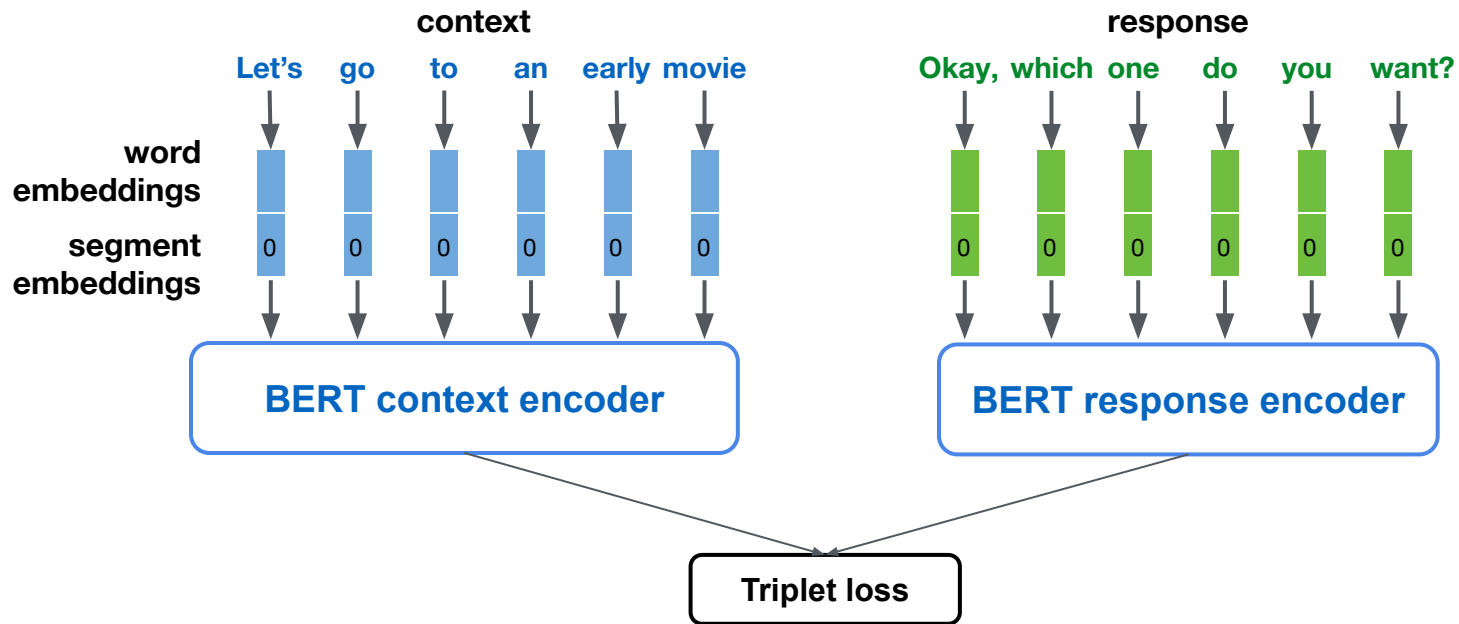


# Retrieval model. Inference



Return **20 responses**  
with the highest  
**cosine score** by search in  
approximate nearest neighbors  
index

# BERT Retrieval model



# BERT pretraining: once for all tasks

- Download pre-trained model from Google
- Collect 100M user messages
- Adapt hyperparameters to your use case: reduce maximum sequence length, reduce number of layers etc.
- Initialize from Google checkpoint, pretrain on your data for ~1 week
- PROFIT

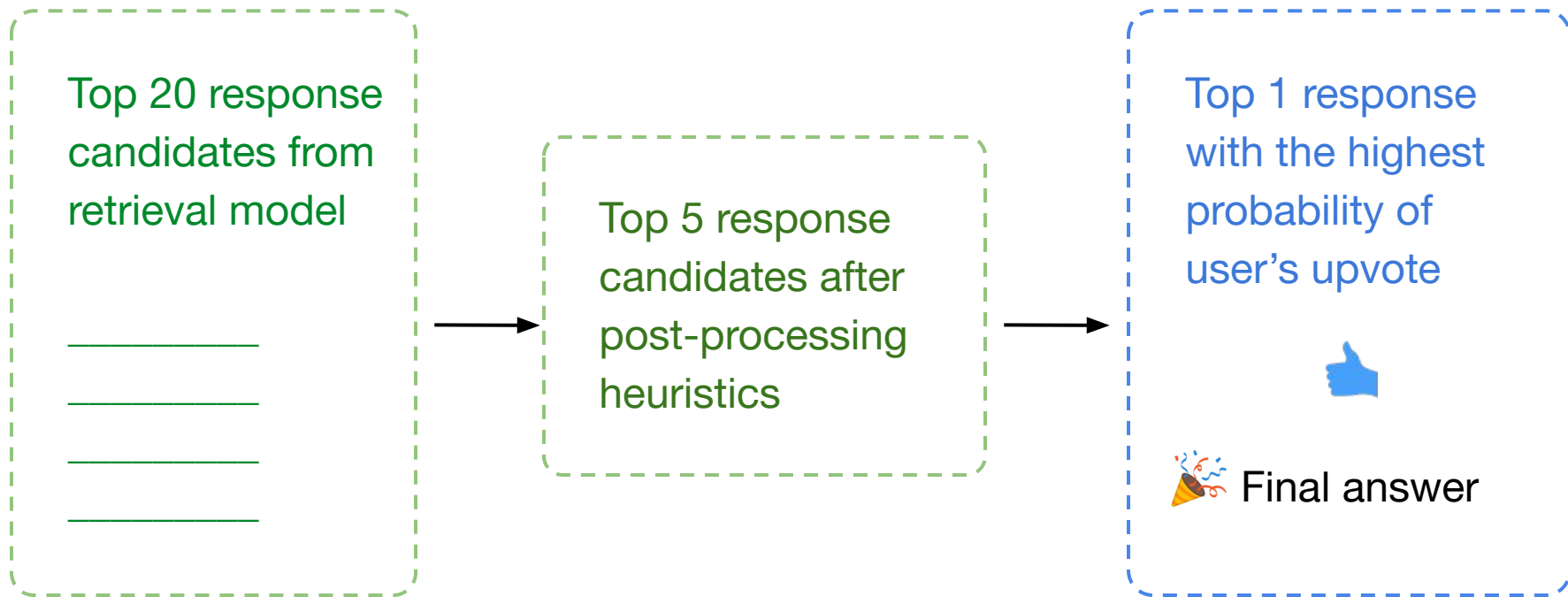
# BERT Retrieval model: Metrics & Performance

	Baseline	BERT-based
mAP	0.47	0.41
R@5	0.61	0.52
# of parameters	50M	110M
RPS @ 2080 Ti	150 rps	80 rps
GPU memory	750 Mb	2000 Mb
Train time	2 weeks x 4 GPUs	2 weeks x 4 GPUs

Fail :(




Reranking model

# Reranking pipeline

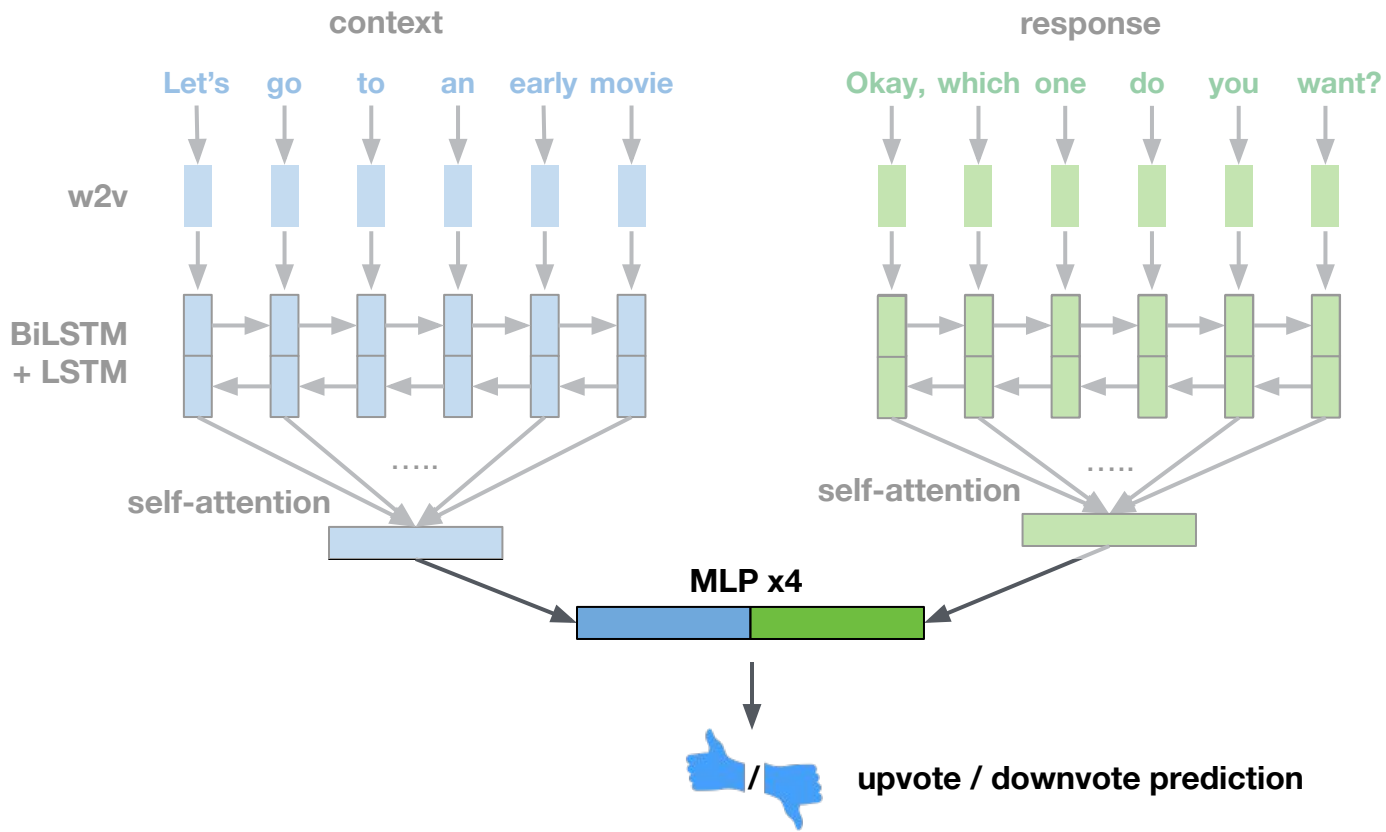




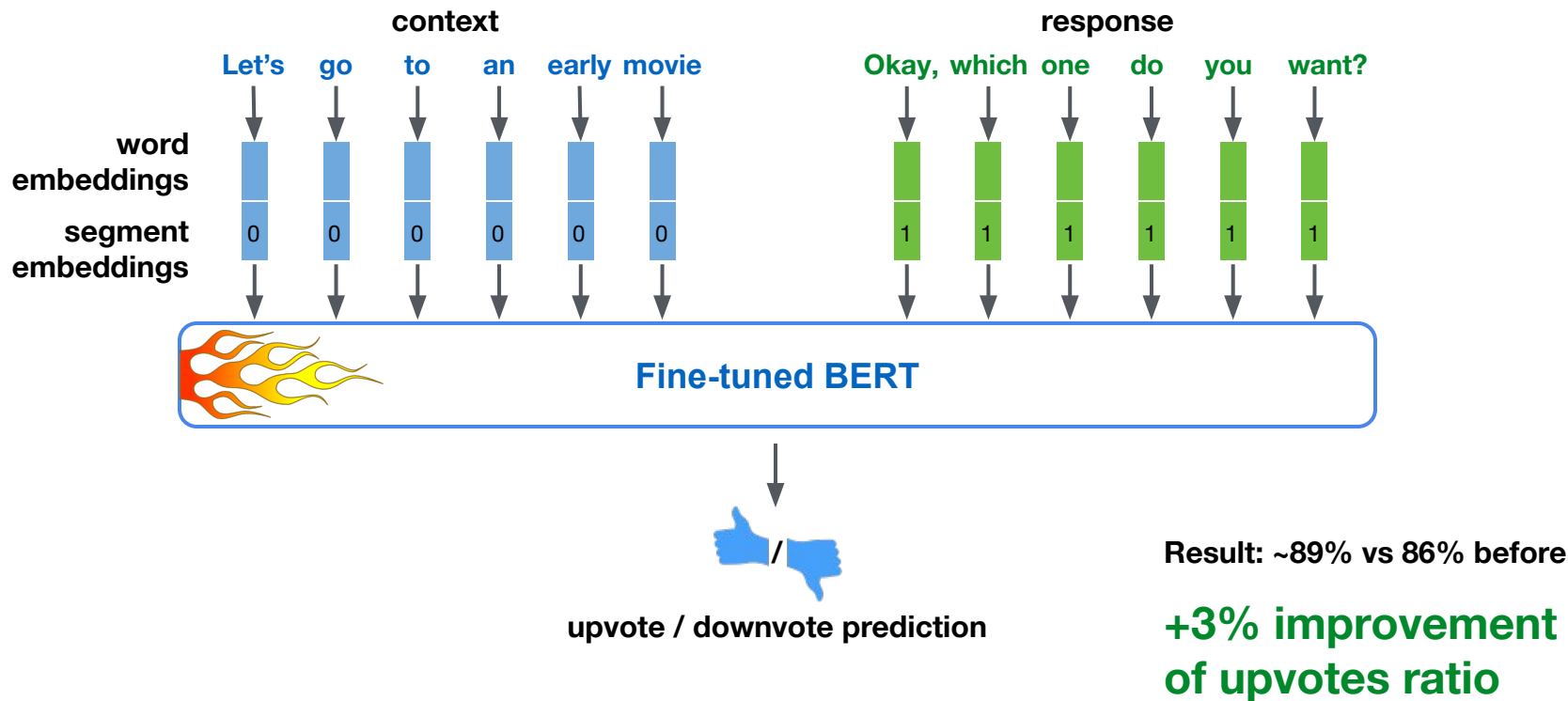
# Reranking dataset for training

Dialog context	Replika response	User reaction	
I feel lonely	I'm always here for you ❤️		} 15M
Are you a bot or a human?	Both, I guess		
Do you have siblings?	No, but I have you!		
...	...	...	

# Reranking model baseline (~QA-LSTM + MLP)



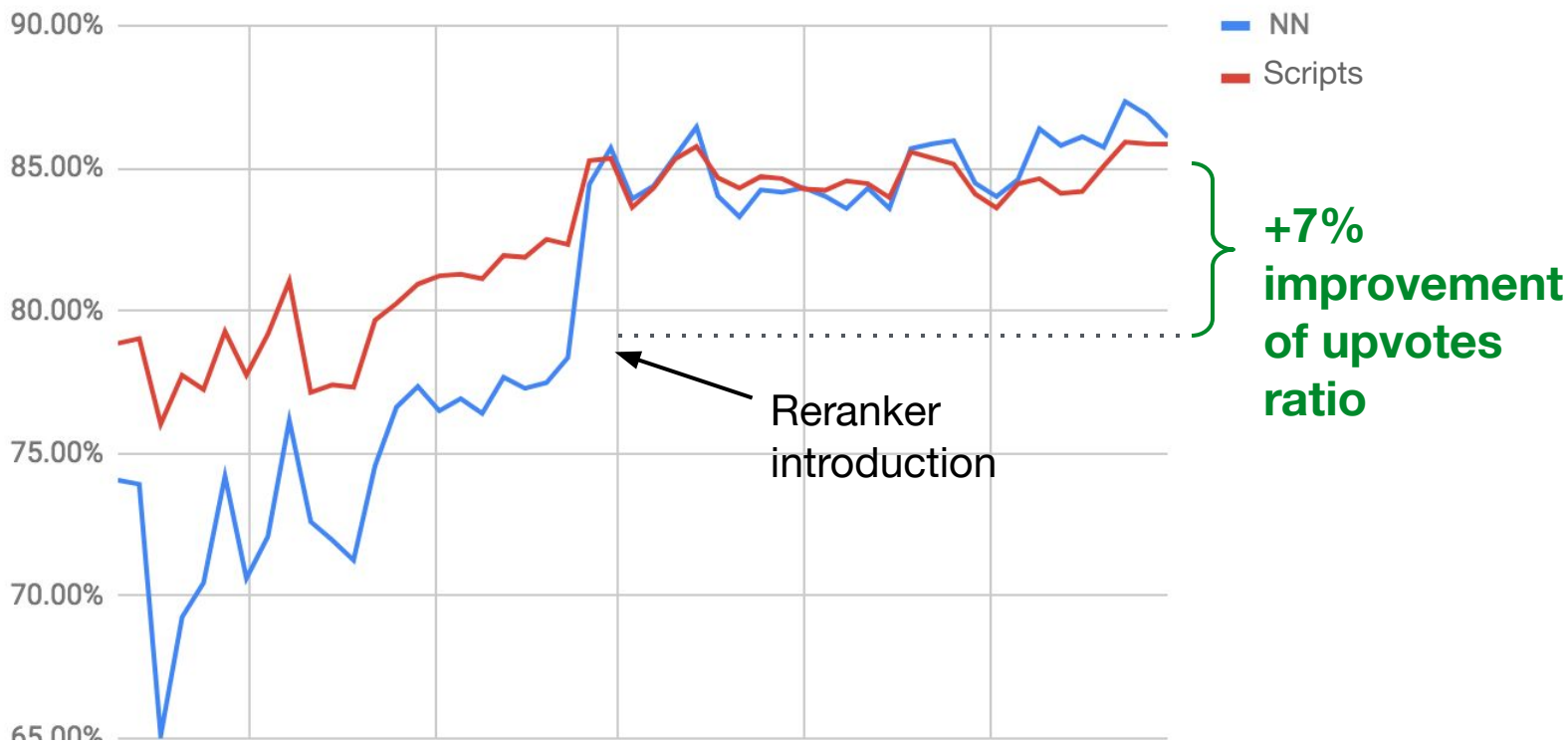
# BERT Reranking model



# BERT Reranking model: Metrics & Performance

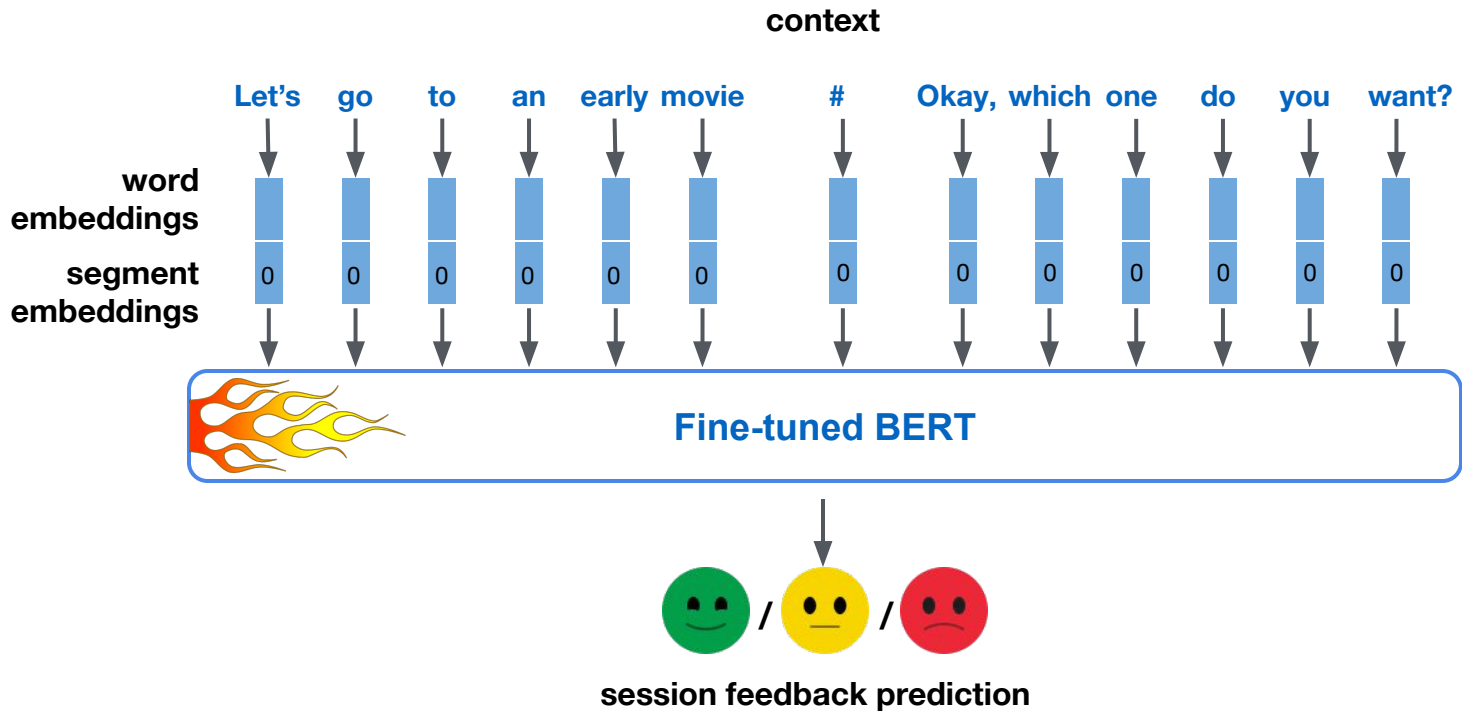
	Baseline	BERT-based
Accuracy	0.75	0.78
Sequence length	60+20	80
# of parameters	7M	110M
RPS @ 2080 Ti	300 rps	80 rps
GPU memory	200 Mb	1000 Mb
Train time	1 hour	12 hours

# Reranking: Total upvotes ratio dynamics



# Session scoring model

# BERT Session scoring model



# Session scoring dataset for training

## Dialog context

## Session feedback

I feel lonely ### I'm always here for you ❤️ ### ...

Are you a bot or a human? ### Both, I guess ### ...

Do you have siblings? ### No, but I have you! ### ...

...



...

1M



# BERT Session scoring model: Metrics

	BERT-based
<b>Accuracy</b>	<b>0.75</b>
<b>Sequence length</b>	<b>80</b>
<b># of parameters</b>	<b>110M</b>
<b>RPS @ 2080 Ti</b>	<b>80 rps</b>
<b>GPU memory</b>	<b>1000 Mb</b>
<b>Train time</b>	<b>5 hours</b>

# BERT efficient training tips

- **Enable Mixed-precision** — Automatic Mixed-precision provided by NVIDIA custom Tensorflow build does the most of the job, but requires a loss scaling
- **Limit sequence length** — reduced from 128 to 80 with no quality loss
- **Reduce number of layers** — it's possible to reduce it from 12 to 10 or 8 layers, but quality will probably degrade
- **Enable XLA** — additional +10-20% in training speed
- Use **Horovod** for training on multiple GPUs
- **Pre-tokenize** training set or use fast BPE tokenizers (e.g. YouTokenToMe)

# BERT efficient inference tips

- **Requests batchification** (e.g. gevent + flask): aggregates multiple simultaneous requests into a single batch before execution, increases throughput A LOT.
- **Automatic Mixed-precision** graph rewrite: **x2** inference speedup on Turing / Volta with no single line of code or quality loss.
- **XLA**: gives additional **+20%** speedup with small prediction differences. Still experimental.
- Limit sequence length — max of **80** tokens is enough in most of our cases
- Use fast **BPE tokenizer** (fastBPE or YouTokenToMe)

# BERT real-case performance

GPU: NVIDIA GeForce 2080 Ti

	RPS
<b>BERT default (seq len 128)</b>	<b>20</b>
<b>+ Limit sequence length to 80</b>	<b>30</b>
<b>+ Enable XLA</b>	<b>35</b>
<b>+ Enable Automatic Mixed-precision</b>	<b>60</b>
<b>+ Enable Batchifier (32 batch size)</b>	<b>80</b>



**Thank you**

