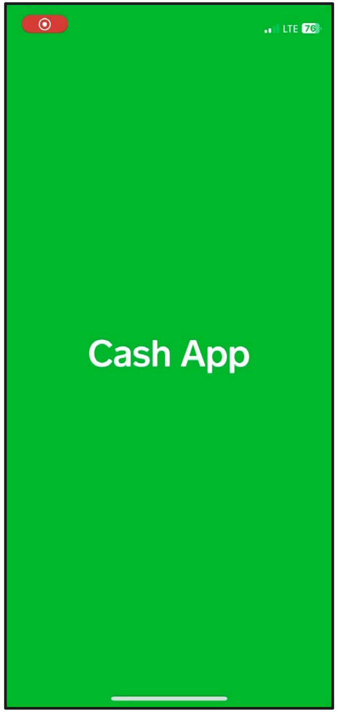
DATA+AI SUMMIT
BY databricks

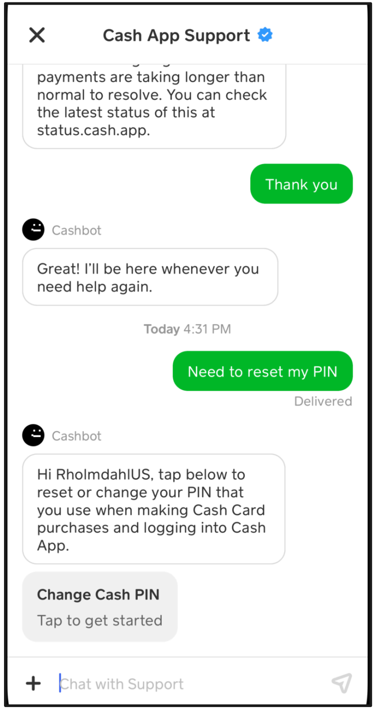
HOW CASH APP TRAINS LARGE LANGUAGE MODELS FOR CUSTOMER SUPPORT

Dean Wyatte

June 11, 2024

CASH APP CUSTOMER SUPPORT





payments are taking longer than normal to resolve. You can check the latest status of this at status.cash.app.

Thank you

Cashbot

Great! I'll be here whenever you need help again.

Today 4:31 PM

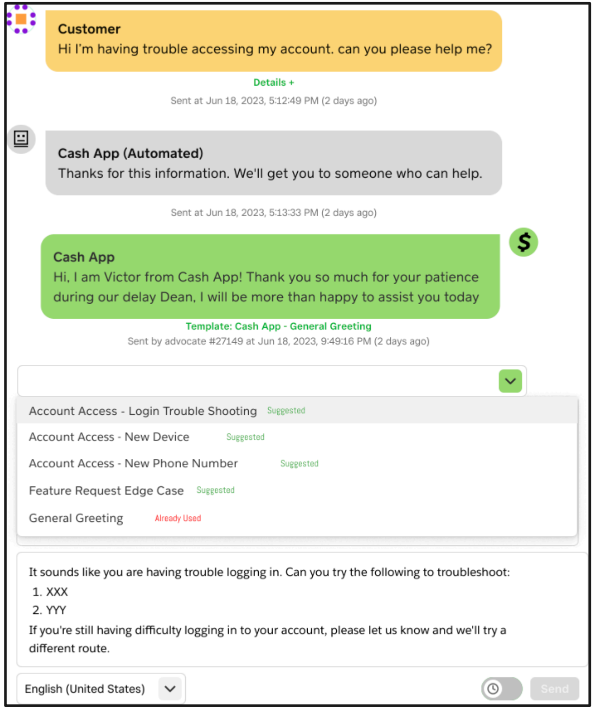
Need to reset my PIN

Delivered

Cashbot

Hi RholmdahlUS, tap below to reset or change your PIN that you use when making Cash Card purchases and logging into Cash App.

Change Cash PIN
Tap to get started



Customer
Hi I'm having trouble accessing my account. can you please help me?

Details +

Sent at Jun 18, 2023, 5:12:49 PM (2 days ago)

Cash App (Automated)
Thanks for this information. We'll get you to someone who can help.

Sent at Jun 18, 2023, 5:13:33 PM (2 days ago)

Cash App
Hi, I am Victor from Cash App! Thank you so much for your patience during our delay Dean, I will be more than happy to assist you today

Template: Cash App - General Greeting

Sent by advocate #27149 at Jun 18, 2023, 9:49:16 PM (2 days ago)

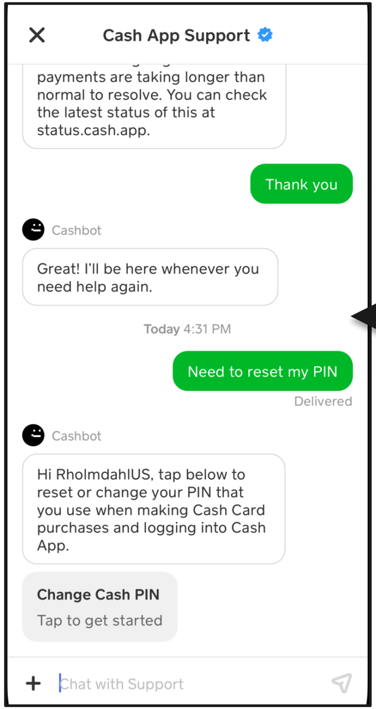
Account Access - Login Trouble Shooting Suggested
Account Access - New Device Suggested
Account Access - New Phone Number Suggested
Feature Request Edge Case Suggested
General Greeting Already Used

It sounds like you are having trouble logging in. Can you try the following to troubleshoot:

1. XXX
2. YYY

If you're still having difficulty logging in to your account, please let us know and we'll try a different route.

CASH APP CUSTOMER SUPPORT



payments are taking longer than normal to resolve. You can check the latest status of this at status.cash.app.

Thank you

Cashbot

Great! I'll be here whenever you need help again.

Today 4:31 PM

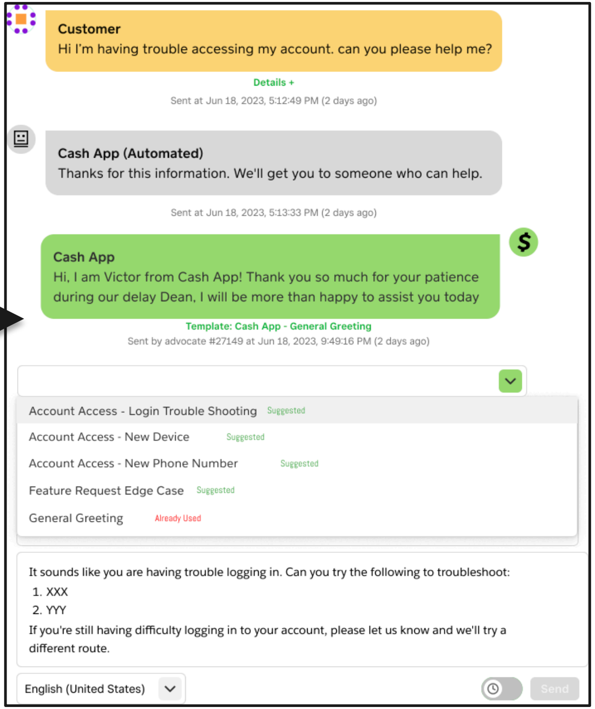
Need to reset my PIN

Delivered

Cashbot

Hi RholmdahlUS, tap below to reset or change your PIN that you use when making Cash Card purchases and logging into Cash App.

LLM



Customer
Hi I'm having trouble accessing my account. can you please help me?

Sent at Jun 18, 2023, 5:12:49 PM (2 days ago)

Cash App (Automated)
Thanks for this information. We'll get you to someone who can help.

Sent at Jun 18, 2023, 5:13:33 PM (2 days ago)

Cash App
Hi, I am Victor from Cash App! Thank you so much for your patience during our delay Dean, I will be more than happy to assist you today

Template: Cash App - General Greeting
Sent by advocate #27149 at Jun 18, 2023, 9:49:16 PM (2 days ago)

Account Access - Login Trouble Shooting Suggested
Account Access - New Device Suggested
Account Access - New Phone Number Suggested
Feature Request Edge Case Suggested
General Greeting Already Used

It sounds like you are having trouble logging in. Can you try the following to troubleshoot:

1. XXX
2. YYY

If you're still having difficulty logging in to your account, please let us know and we'll try a different route.

CUSTOMER SUPPORT IS A CLOSED DOMAIN

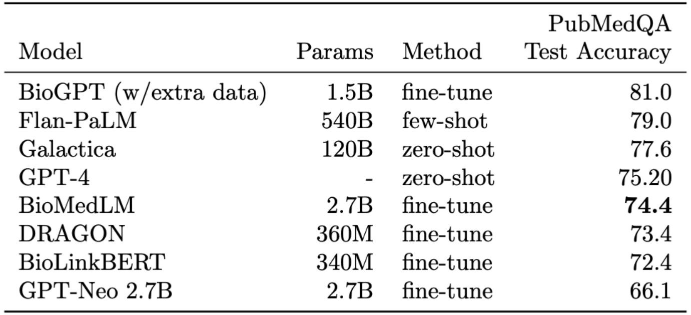
- Typical LLMs like OpenAI’s GPT family and Meta’s Llama are open-domain assistants
  - Knowledgeable about many topics
  - Can be instructed to perform many tasks

- Customer support is a closed domain
  - Assistants only need knowledge about their domain (Cash App, general consumer finance)
  - Assistants should only perform tasks related to customer support (don’t code, don’t write poetry)

- Closed domains allow specialization
  - Improved control over model size and latency
  - Models less likely to be jailbroken to perform arbitrary tasks
  - Running models in-house improves privacy, PII may even be required for some domains / tasks

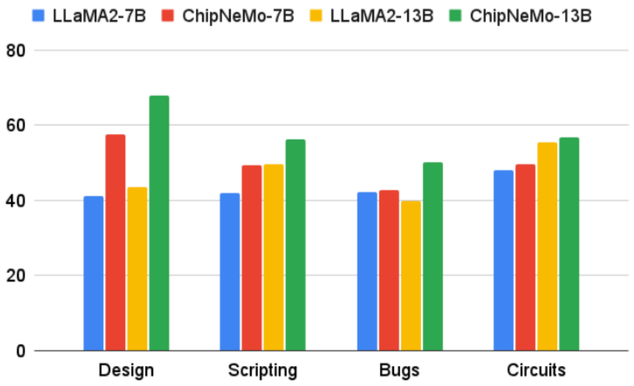
LLMS IN CLOSED DOMAINS

- BioMedLM (Bolton et al., 2022)
  - 2.7B params
  - 35B tokens from The Pile filtered to biomedical literature
- ChipNeMo (Liu et al., 2023)
  - 7B and 13B params
  - 23B tokens of chip design docs / code + 128K instruction tokens
- Code Llama (Rozière et al., 2023)
  - 7B, 13B, 34B, and 70B params
  - 500B-1T tokens depending on model size



Model	Params	Method	PubMedQA Test Accuracy
BioGPT (w/extra data)	1.5B	fine-tune	81.0
Flan-PaLM	540B	few-shot	79.0
Galactica	120B	zero-shot	77.6
GPT-4	-	zero-shot	75.20
BioMedLM	2.7B	fine-tune	**74.4**
DRAGON	360M	fine-tune	73.4
BioLinkBERT	340M	fine-tune	72.4
GPT-Neo 2.7B	2.7B	fine-tune	66.1

Table 10: PubMedQA Performance of Various Systems



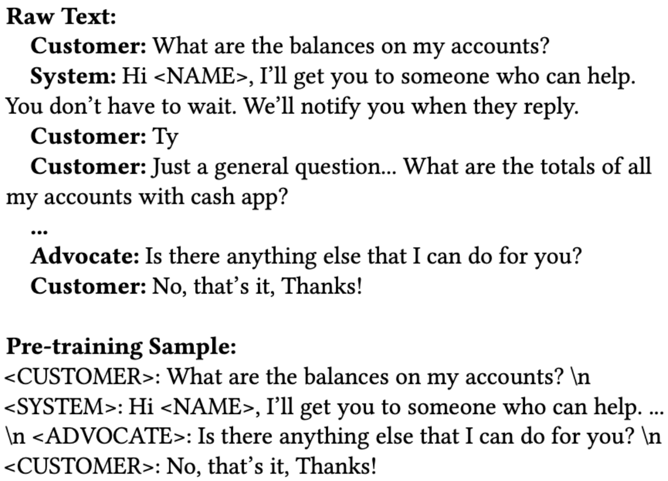
(a) Chip Design Domain Benchmarks.

PRE-TRAINING LLMS FOR CUSTOMER SUPPORT

PRE-TRAINING LLMS FOR CUSTOMER SUPPORT

- Start with the simplest pre-training
    - 10-100B tokens of raw transcripts
    - Hallucinations possible (hidden information)

- Typical tools, primary differentiator is efficiency
    - Hugging Face transformers
    - Microsoft DeepSpeed / PyTorch FSDP
    - NVIDIA Megatron-LM
    - Third-party APIs



Raw Text:

Customer: What are the balances on my accounts?

System: Hi <NAME>, I'll get you to someone who can help. You don't have to wait. We'll notify you when they reply.

Customer: Ty

Customer: Just a general question... What are the totals of all my accounts with cash app?

...

Advocate: Is there anything else that I can do for you?

Customer: No, that's it, Thanks!

Pre-training Sample:

<CUSTOMER>: What are the balances on my accounts? \n <SYSTEM>: Hi <NAME>, I'll get you to someone who can help. ... \n <ADVOCATE>: Is there anything else that I can do for you? \n <CUSTOMER>: No, that's it, Thanks!

$$\sum_{i=1}^{N} log P(x_i | x_1 ... x_{i-1})$$

PRE-TRAINING LLMS FOR CUSTOMER SUPPORT

Generated data exhibits Customer Support knowledge, BOLD is prompt

<CUSTOMER>: What are the balances on my accounts?
<SYSTEM>: Hi <NAME>, I’ll get you to someone who can help. This may take a few hours, but you don't have to wait. We'll notify you when they reply.
<ADVOCATE>: Hey there, this is <NAME> with Cash App, I apologize by our delay replying back, now I am here to assist you, just give me a few more time to check out your account.
<ADVOCATE>: I understand your concern and I would be more than glad to help.
<ADVOCATE>: To view your balance:

- Sign into the app
- You'll see your balance listed in the Money tab at the bottom left of the home screen

<ADVOCATE>: If there is anything else I can do for you, just let me know and I will be thrilled to help.
<ADVOCATE>: It was a pleasure assisting you today. If you have any further questions, feel free to reach back out - we'd love to help!

PRE-TRAINING LLMS FOR CUSTOMER SUPPORT

Generated data is self-consistent, BOLD is prompt

…
<CUSTOMER>: That didn’t help and now I’m mad!
<SYSTEM>: "How would you rate our interaction?”
<CUSTOMER>: 😠
<SYSTEM>: Thanks for your feedback!

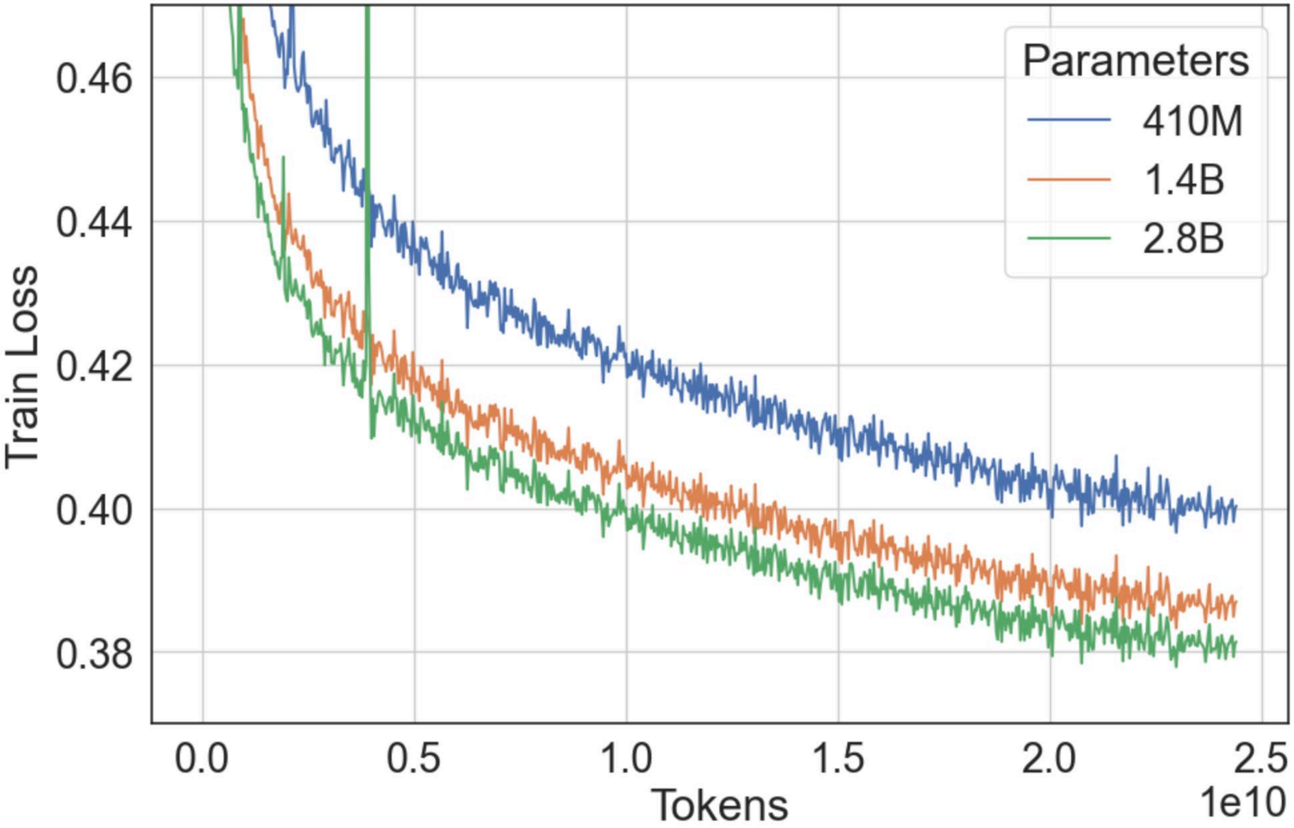
…
<CUSTOMER>: That solved everything!
<SYSTEM>: How would you rate our interaction?
<CUSTOMER>: 😍
<SYSTEM>: Thanks for your feedback!
<CUSTOMER>: Thanks!
<ADVOCATE>: You’re very welcome! Please feel free to reach back out if anything else comes up. Have a great day!
<CUSTOMER>: Thank you! You as well!

PRE-TRAINING LLMS FOR CUSTOMER SUPPORT

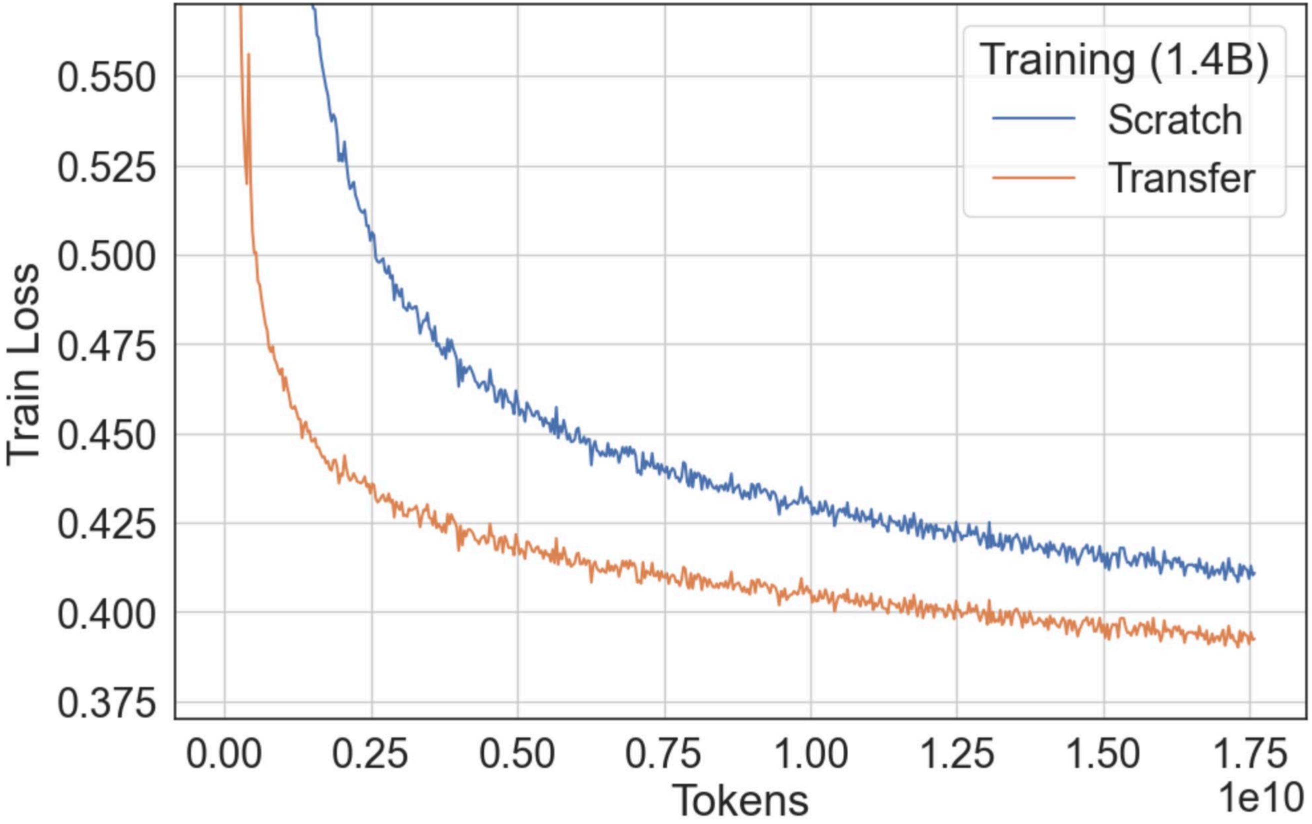
Generated data contains hallucination, BOLD is prompt

<CUSTOMER>: What is this $9.99 transaction every month? I don't authorize that.
<SYSTEM>: Hi <NAME>, advocates will be back online in the morning. We'll notify you when they reply.
<CUSTOMER>: Thank you.
<ADVOCATE>: Hey <NAME>, <NAME> with Cash App Support here! Sorry for the delay! I am here to assist in any way I can.
<ADVOCATE>: The 9.99 payment from GOOGLE*YOUTUBE MUSIC is a pre-approved recurring subscription payment. A merchant can force an authorization on your card and continue to pull funds from your balance, even if you've disabled or reported the card. You'll need to reach out and ask them to remove your card from the system and issue a refund.
<ADVOCATE>: Do you have any more questions or concerns I can assist you with?
<ADVOCATE>: If you need more help, feel free to reach back out during business hours.
<SYSTEM>: How would you rate our interaction?
<CUSTOMER>: 🙂
<SYSTEM>: Thanks for your feedback!

PRE-TRAINING LLMS FOR CUSTOMER SUPPORT

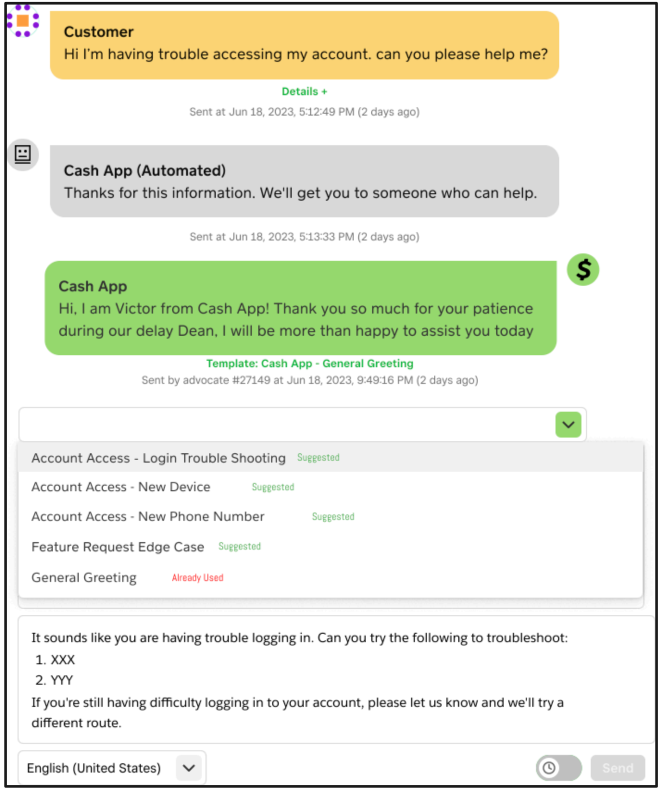


PRE-TRAINING LLMS FOR CUSTOMER SUPPORT



FINE-TUNING LLMS FOR CLASSIFICATION

FINE-TUNING LLMS FOR CLASSIFICATION



Customer
Hi I'm having trouble accessing my account. can you please help me?

Details +
Sent at Jun 18, 2023, 5:12:49 PM (2 days ago)

Cash App (Automated)
Thanks for this information. We'll get you to someone who can help.

Sent at Jun 18, 2023, 5:13:33 PM (2 days ago)

Cash App
Hi, I am Victor from Cash App! Thank you so much for your patience during our delay Dean, I will be more than happy to assist you today

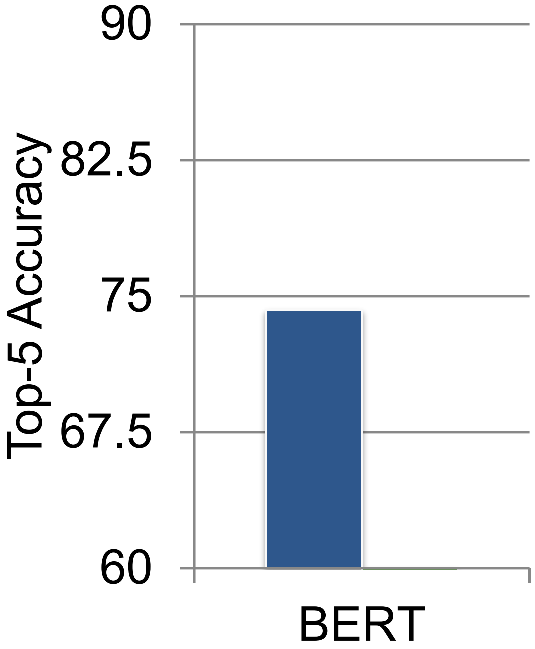
Template: Cash App - General Greeting
Sent by advocate #27149 at Jun 18, 2023, 9:49:16 PM (2 days ago)

Account Access - Login Trouble Shooting Suggested
Account Access - New Device Suggested
Account Access - New Phone Number Suggested
Feature Request Edge Case Suggested
General Greeting Already Used

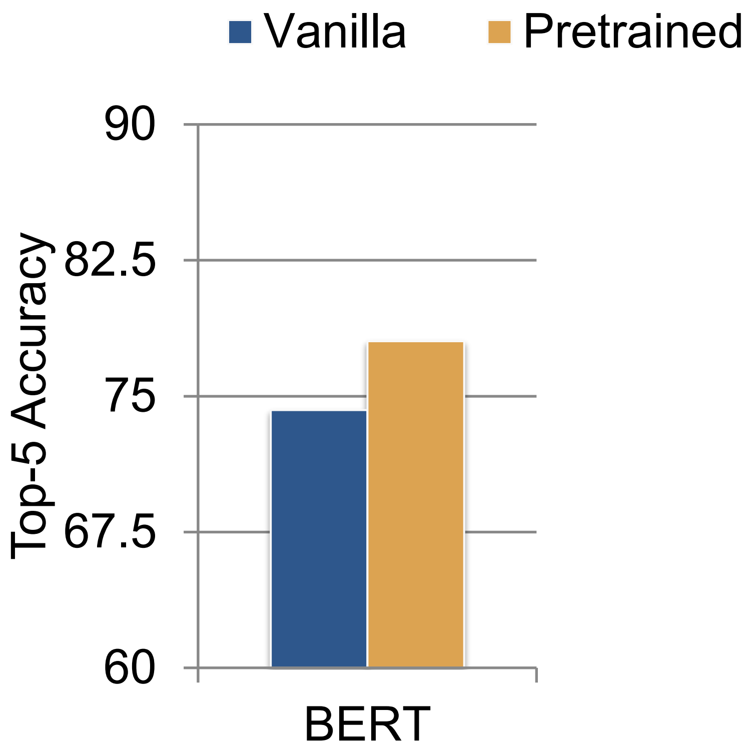
It sounds like you are having trouble logging in. Can you try the following to troubleshoot:
1. XXX
2. YYY
If you're still having difficulty logging in to your account, please let us know and we'll try a different route.

- Use Case: Top-K classification over 100s-1000s of template responses
- BERT / encoders or few-shot LLMs are the standard approach
  - BERT / encoders remain relatively unscaled since LLMs popularized
  - It's not clear how to few-shot over thousands of classes
- Can we treat classes as tokens and generate one new token (masked to class tokens)?
  - Benefit from LLM scale without hallucination

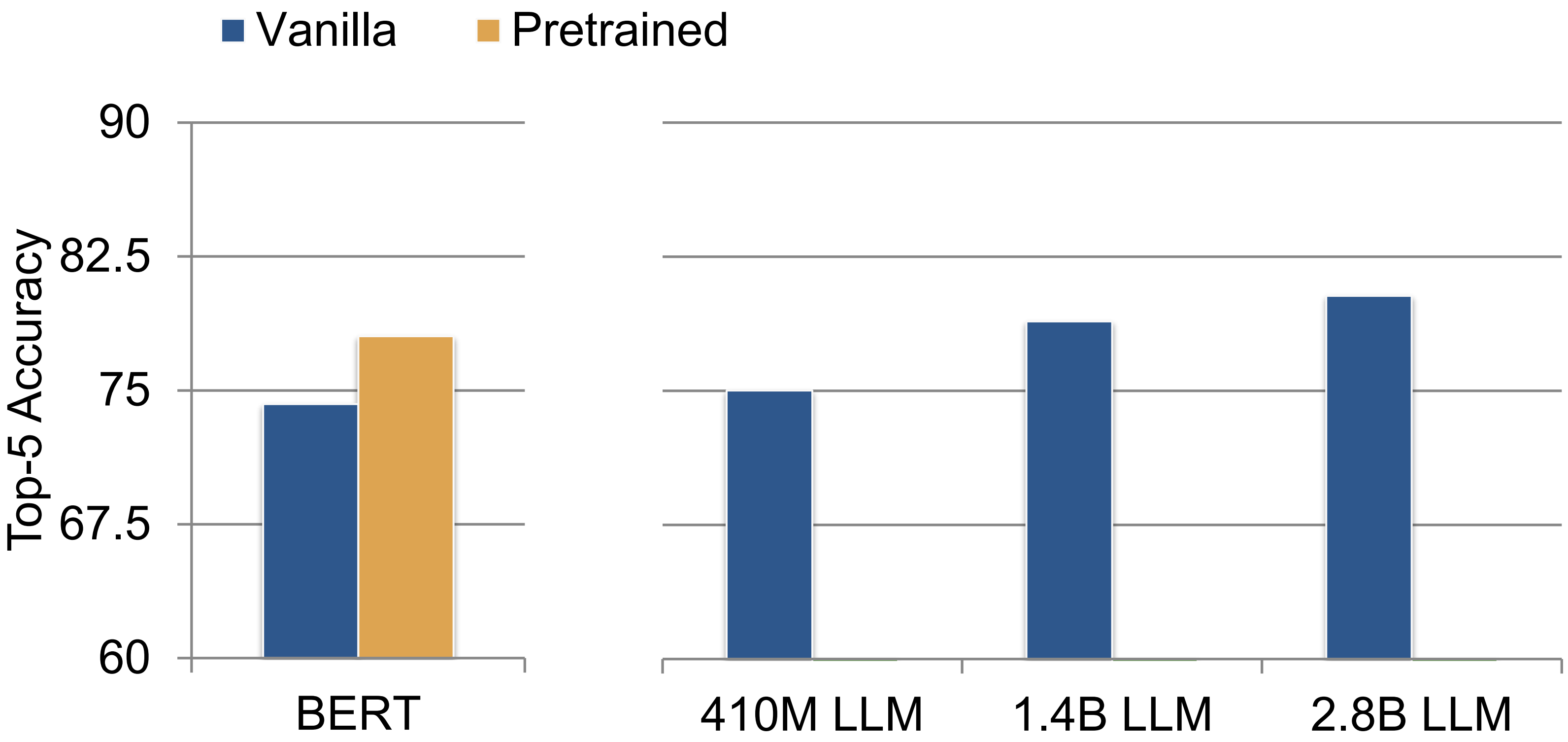
FINE-TUNING LLMS FOR CLASSIFICATION



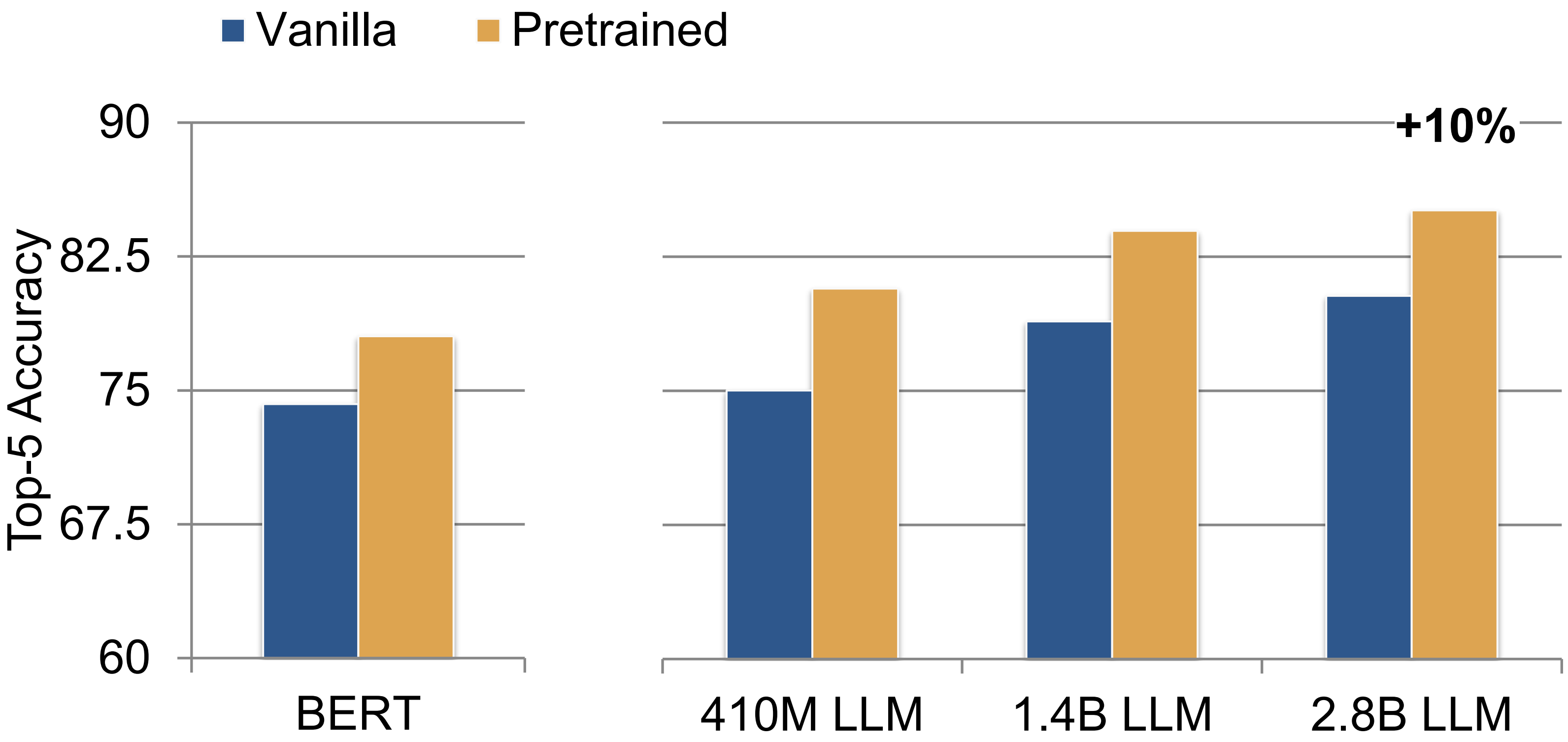
FINE-TUNING LLMS FOR CLASSIFICATION



FINE-TUNING LLMS FOR CLASSIFICATION

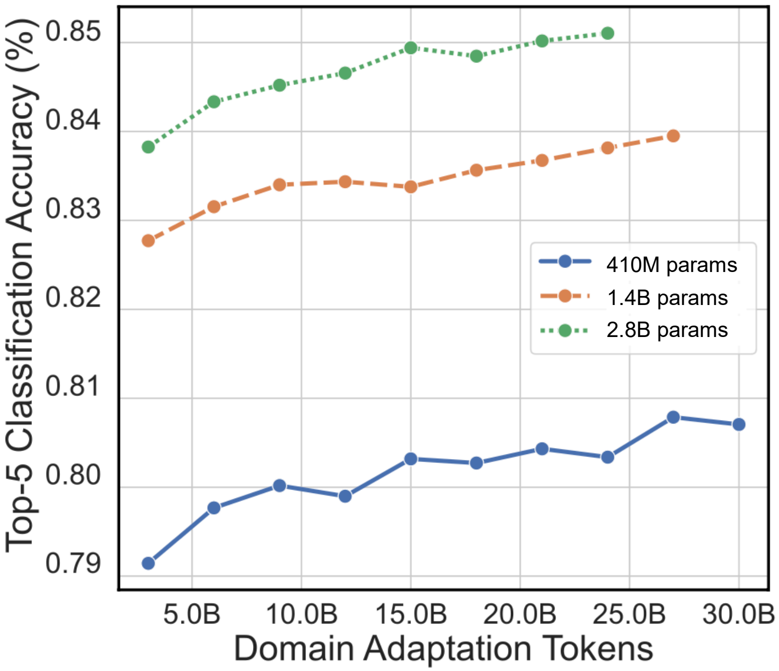


FINE-TUNING LLMS FOR CLASSIFICATION

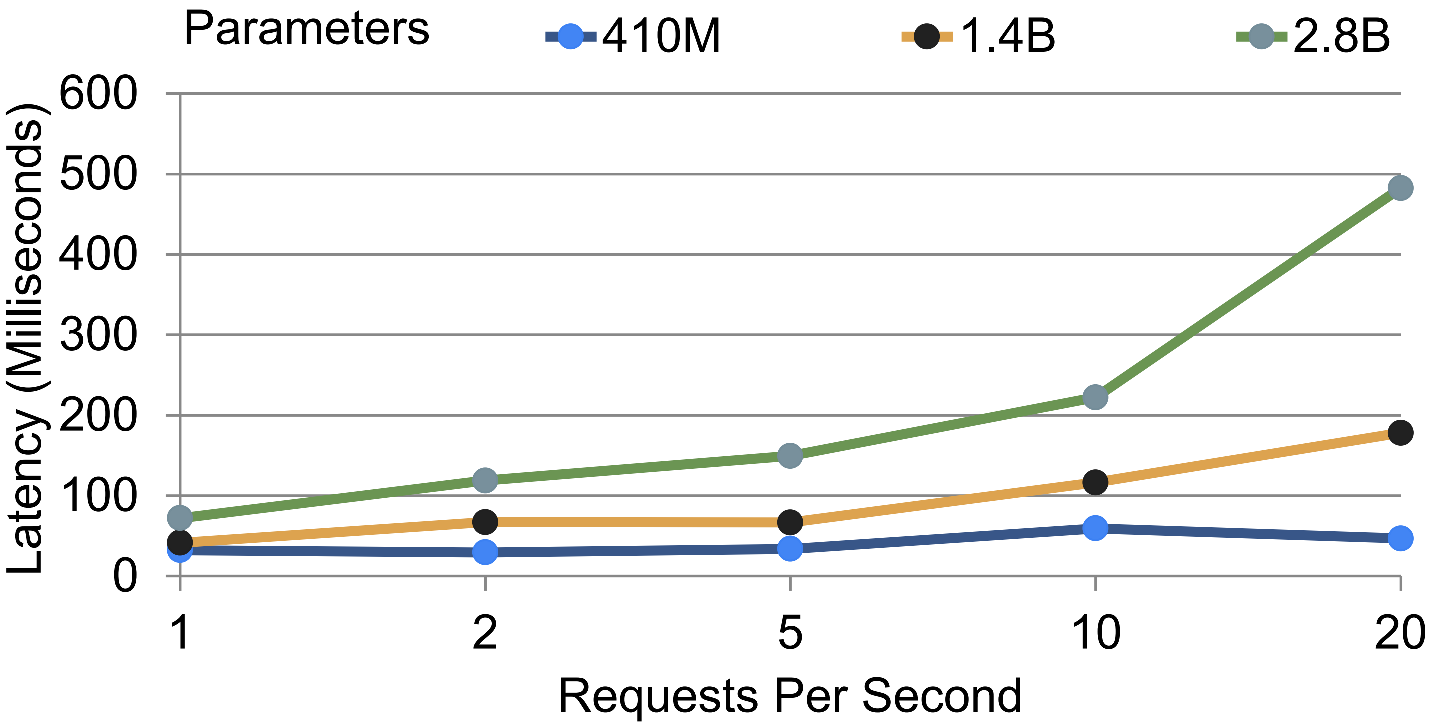


FINE-TUNING LLMS FOR CLASSIFICATION

- Effect of model size
  - Not yet saturated, but diminishing returns
- Effect of number of tokens.
  If starting with off-the-shelf model
  - First 3B tokens most important (+3% acc.)
  - Linear effect after (+1.5% acc. over 27B tokens)



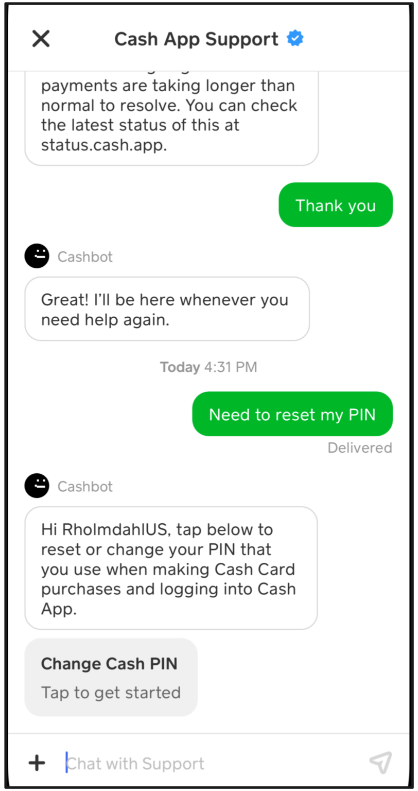
FINE-TUNING LLMS FOR CLASSIFICATION



NVIDIA TensorRT 1xA10G. Sequence lengths sampled from Cash App data

FINE-TUNING LLMS FOR GENERATION

FINE-TUNING LLMS FOR GENERATION



payments are taking longer than normal to resolve. You can check the latest status of this at status.cash.app.

Thank you

Cashbot

Great! I'll be here whenever you need help again.

Today 4:31 PM

Need to reset my PIN

Delivered

Cashbot

Hi RholmdahlUS, tap below to reset or change your PIN that you use when making Cash Card purchases and logging into Cash App.

Change Cash PIN
Tap to get started

- Use Case: Generate appropriate response given partial transcript
- Up until now, we have treated this as classification
  - Model classifies customer intent
  - Rule engine sends mostly hardcoded responses
- LLMs may provide increased flexibility
  - Classification has high maintenance cost, decreasing ROI
  - Generation can tailor response given context

FINE-TUNING LLMS FOR GENERATION

- We treat response generation as a fine-tuning problem
- Some information is hidden and needs retrieved during inference to ground generation
  - Transaction details relevant to the support case
  - Conditioning generation on discrete classification is a powerful tool

Hypothetical example

### TRANSCRIPT
<CUSTOMER>: What is this transaction every month? I don't authorize that.

### TRANSACTION
Customer Name: {Dean}
Transaction Recipient: {Music Subscription Company}
Transaction Amount: {$9.99}
Transaction State: {PAID}

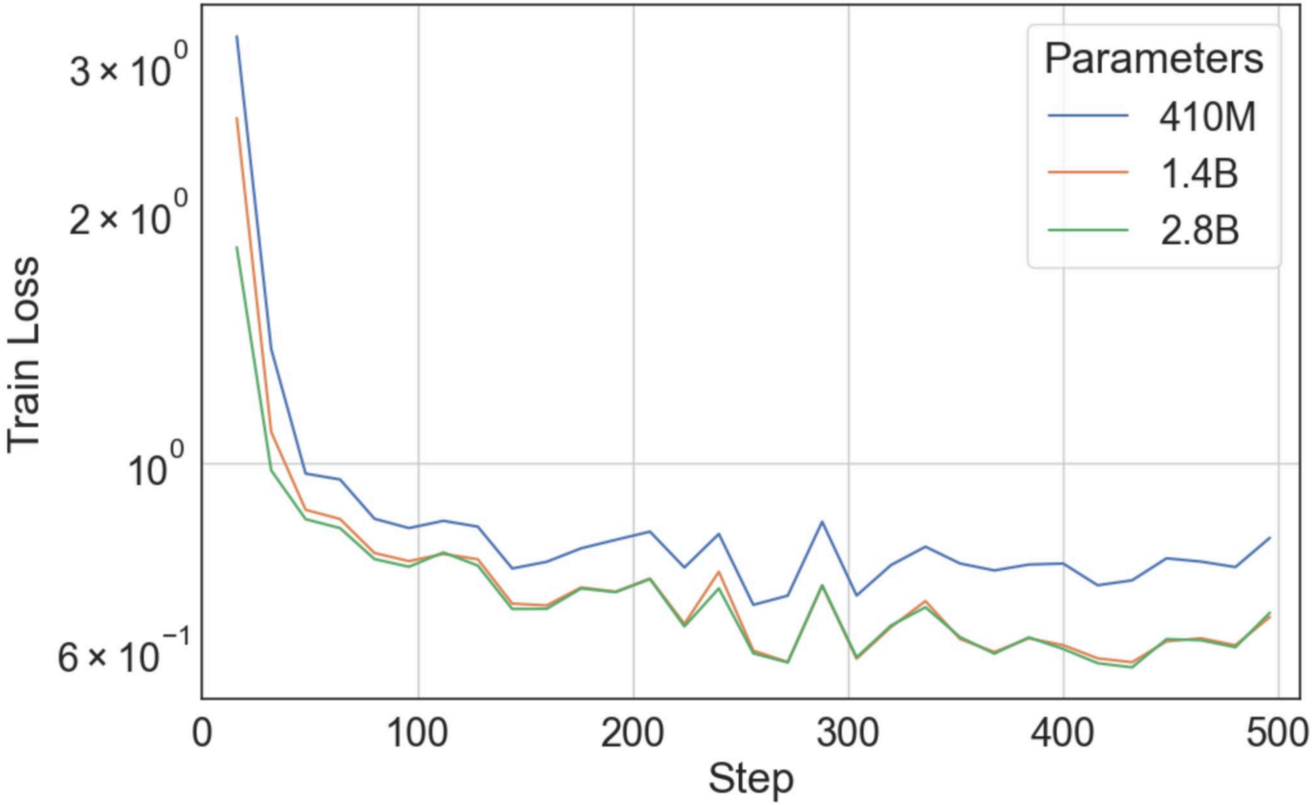
```
{
   “classification”:   “CANCEL_SUBSCRIPTION”,
   “response”:         “Hi Dean, the $9.99 payment from Music
                        Subscription Company is a pre-approved recurring
                        subscription payment. A merchant can force an
                        authorization on your card and continue to pull
                        funds from your balance, even if you've disabled or
                        reported the card. You’ll need to reach out and ask
                        them to remove your card from the system and
                        issue a refund.
}
```

FINE-TUNING LLMS FOR GENERATION

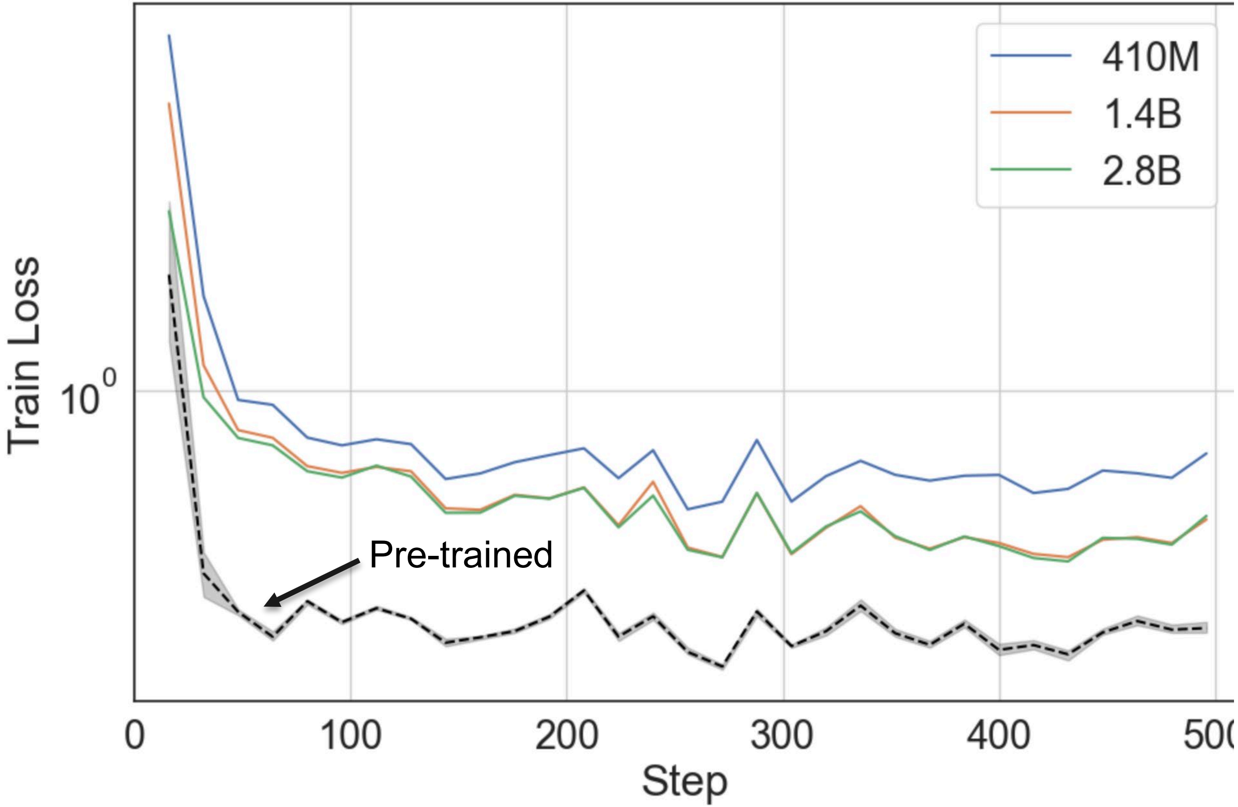
- Evaluation is increasingly important with LLMs
  - With classification we can look at precision / recall, but response quality requires more careful judgement
- Once evaluation is defined pre-training, fine-tuning, retrieval augmentation, and other approaches become an implementation detail with tradeoffs
  - Evaluation accuracy
  - End-to-end latency
  - System complexity
  - Maintenance burden

Criteria	Score
Politeness	100
Unbiased	100
Spelling, grammar, etc.	100
Upholds company policy	100
Realistic	100
Factuality	99.42
Addresses customer issue	97.78
On topic	99.95
…	…

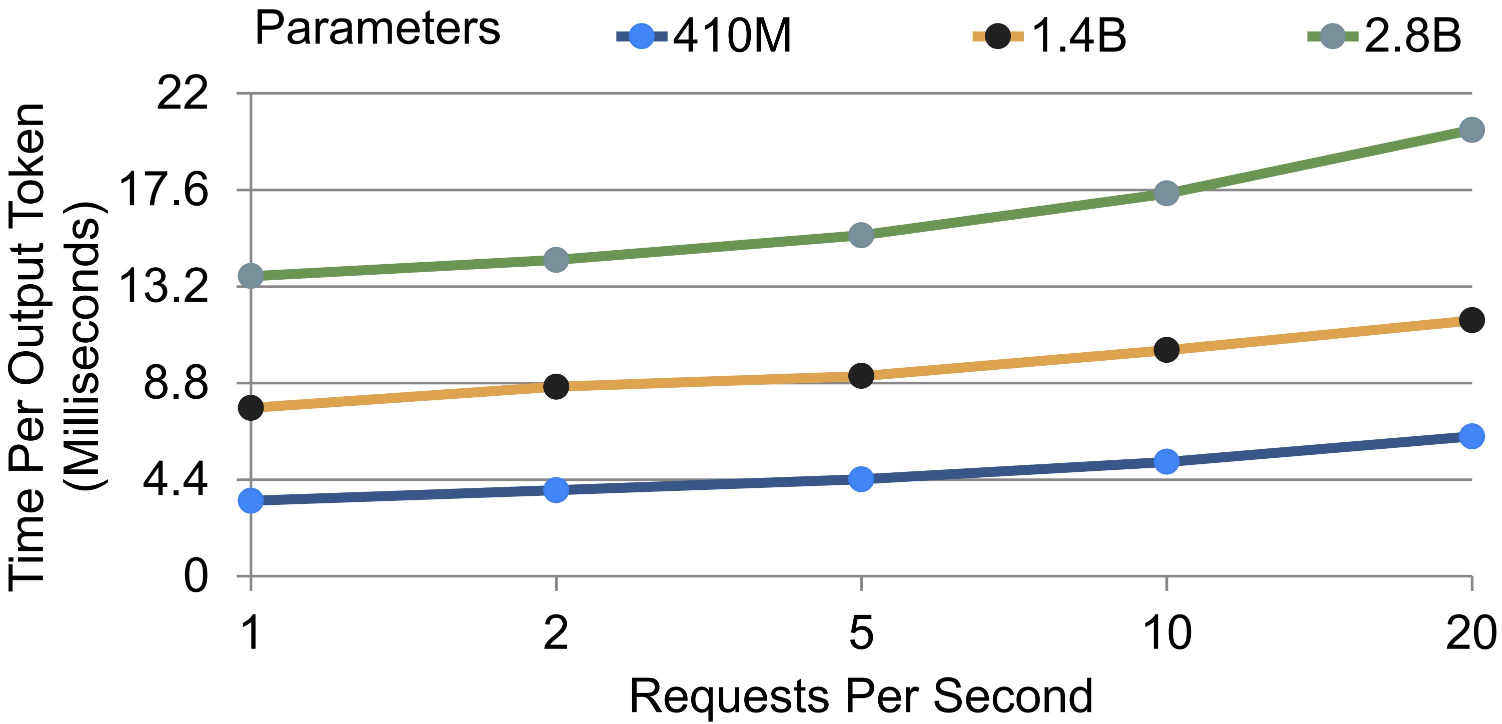
FINE-TUNING LLMS FOR GENERATION



FINE-TUNING LLMS FOR GENERATION



FINE-TUNING LLMS FOR GENERATION



NVIDIA 1xA10G. Prompt length < 10 tokens

SUMMARY

- Customer support is a closed domain and we can use this to our advantage to develop specialized LLMs

- We still require classification throughout our systems, but training LLMs (even small ones) is an attractive evolution to encoders or off-the-shelf LLMs

- Everything is an LLM: A single model architecture / training objective for classification and generation simplifies model development

DATA+AI SUMMIT