LLMs for Email Classification @ Deutsche Bahn

October 13, 2023

Chia-Jung Chang, Nils Marthiensen, Neelesh Bhalla, Kirtesh Patel

Management summary

The task

- Email content analysis and categorization.
- Annotation tool development.
- LLM fine-tuning with annotated emails.
- Testing on a selected email sample.
- Results evaluation using test set.
- Potential improvement strategies.

The dataset

- 158 employees' pre-crash emails.
- ~500k raw text files.
- Unstructured data with various content.
- ~1700 emails manually labeled with 4-layer system by UC Berkeley students and ~12000 AI-labeled samples.
- Multiple categories per email in the UCB dataset.

The findings

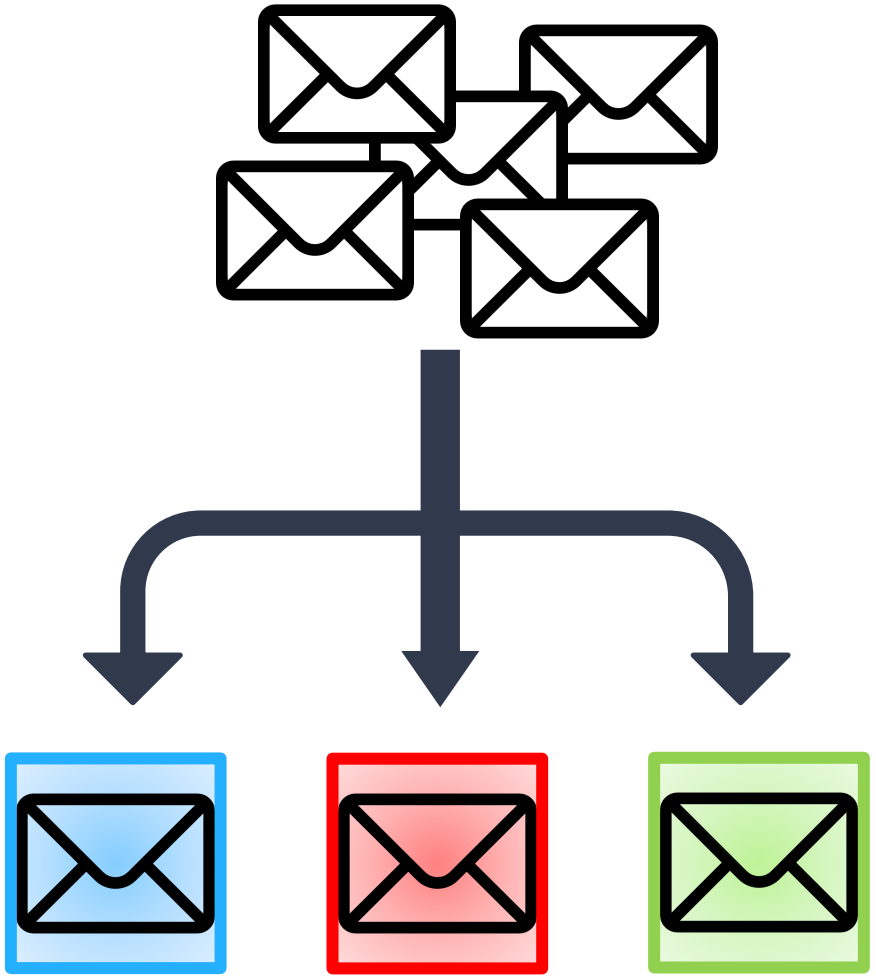
- Manual annotation is expensive and can be replaced by AI methods at lower cost and higher accuracy.
- Categorization accuracy is hard to measure, as several categories are often applicable to one email (At least for the Enron example).
- A lot of data cleaning can be required if the structure of emails changes in any way.
- Fine-tuning models is computationally expensive and skilled talent is rare in the labor market, but improvements are possible even compared to state-of-art LLM-models.

1. The task

2. The dataset

3. The findings

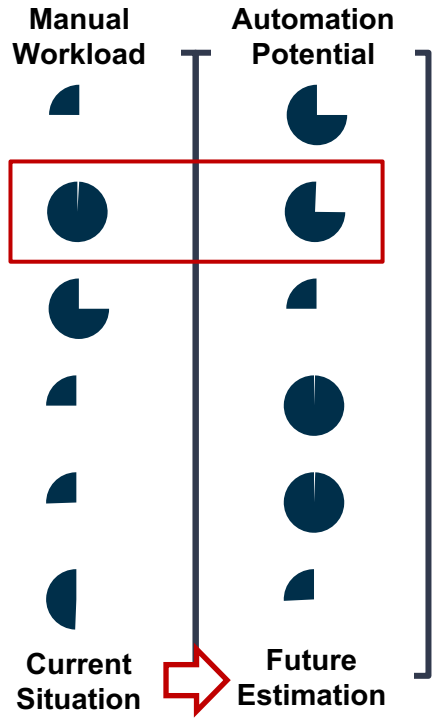
1. The task



General process of LLM-categorization of emails

Assigned Task

1. Small-scale analysis of the email contents and definition of suitable categories.
2. Implementation of an annotation tool to label a subset of the emails with the pre-defined categories.
3. Process of fine-tuning pre-trained LLM-architectures with the annotated emails.
4. Selection of a test sample and inferencing on the fine-tuned LLM.
5. Evaluation of the results using the annotations to get a measure for the quality of the outputs.
6. Improvement of the results by either modifying or changing the LLM and/or increasing the training sample.

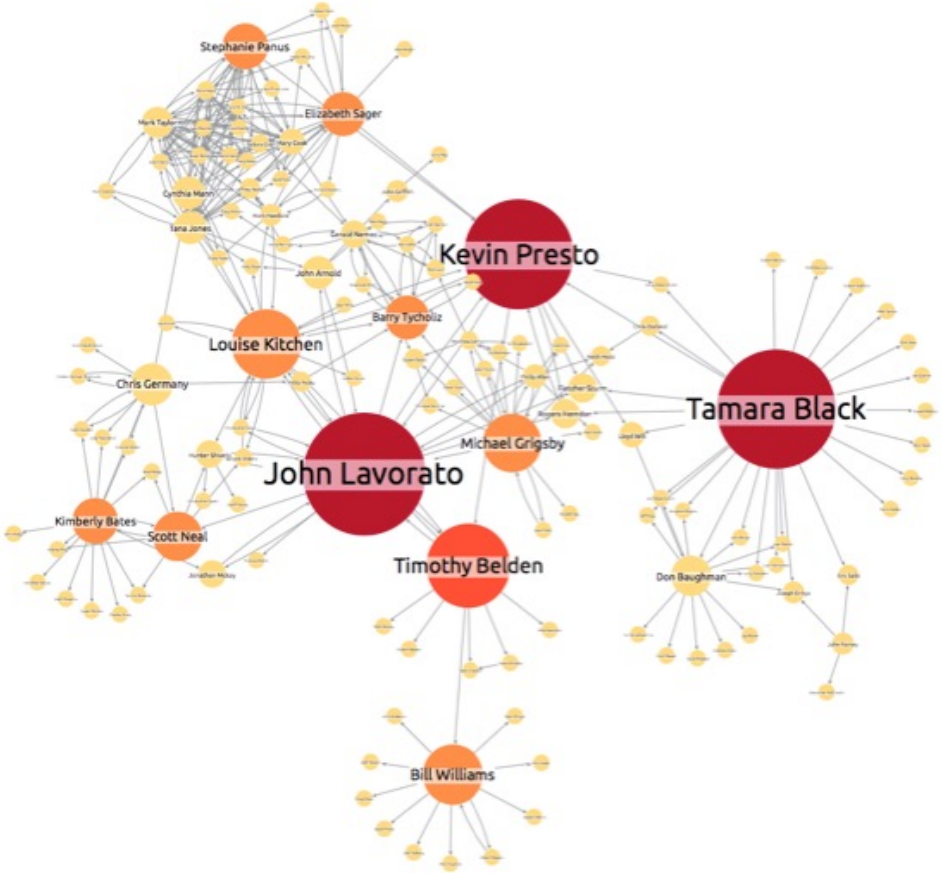


Background and our approach

- Deutsche Bahn receives a lot of emails every day. These concern a wide variety of subjects. To be able to reliably assign categories to the incoming emails would greatly enhance processing and further steps.
- Unfortunately, the data is confidential as it often contains sensitive customer data.
- Due to strict GDPR laws in the EU it is not suitable for a course-project without NDA.

- We worked on a PoC using a different, publicly available dataset – ENRON email Corpus. It addresses the Enron Cooperate scandal and contains raw emails of several persons involved.
- Using this dataset, we were able to conduct the aforementioned steps and can generate valuable insights into the opportunities and risks when conducting such a project.
- Following our experimental research, the groundwork for an internal implementation at Deutsche Bahn is laid.

2. The dataset



Data sources

Public Enron Email Dataset

- Emails of 158 employees before the crash.
- Folders containing ~500k raw text files.
- Largely unstructured: Several reply threads, empty emails or attachments, order of the elements in one text file (ID, sender, receiver, subject, body).
- Wide range of topics, from scandal-related, to standard internal and external communications as well as private conversations.

UC Berkeley Category Labels

- Subset of ~1700 emails manually labelled using a standardized category system for the Enron dataset. Performed by UC Berkeley students as a course project.
- The labelling system involves 4 layers: Coarse genre, included/forwarded information, primary topics, emotional tone.
- Each layer has 8 to 13 subcategories
- One-to-many: One email can have several categories assigned to it.

AI-Generated Labels

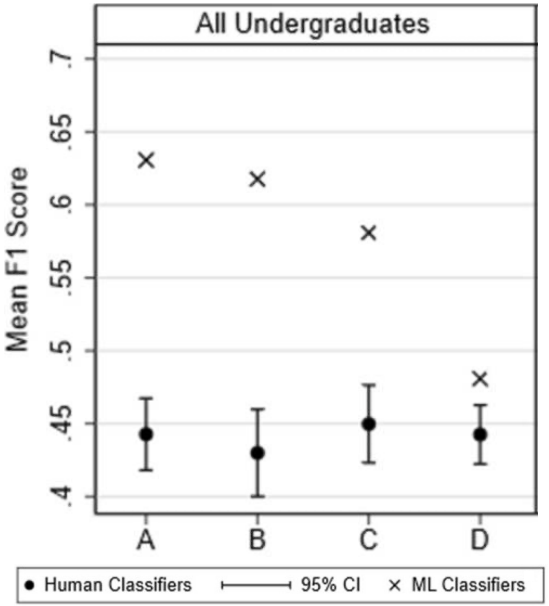
- Subset of ~12000 emails labeled using ChatGPT 3.5 via the API.
- One-to-one: One email gets assigned to one category.
- More training data significantly improved the performance of the fine-tuned models.

Data labeling

Method	Amount classified	Literature F1-Score*	Error reduction measures	Minimum Cost per annotation†
Manual	200	0.43 - 0.45	Two-person-rule	EUR 0.18 *(Total EUR 36)*
General LLM (ChatGPT-3.5)	12000	0.48 - 0.64	None	EUR 0.0005 *(Total EUR 6)*
UC Berkeley Students	1700	/	Two-person-rule	EUR 0.18 *(Total EUR 306)*

*See Goh YC, Cai XQ, Theseira W, Ko G, Khor KA for more details.
†Calculated using german minimum wage of EUR 12 per hour; Sum of time for 50 classifications per group member; UCB time estimate derived from our time methodology.

Related study on classifying abstracts*



Data categories

The original UCB System had to be broken down for this pilot, but more categories can be technically be added.

1 Coarse genre (8)

1.1 Company Business, Strategy, etc. (elaborate in Section 3 [Topics])
1.2 Purely Personal
1.3 Personal but in professional context (e.g., it was good working with you)
1.4 Logistic Arrangements (meeting scheduling, technical support, etc)
1.5 ...

2 Included/forwarded information (13)

2.1 Includes new text in addition to forwarded material
2.2 Forwarded email(s) including replies
2.3 Business letter(s) / document(s)
2.4 News article(s)
2.5 Government / academic report(s)
2.6 Government action(s)
2.7 ...

3 Primary topics (if coarse genre 1.1 is selected) (13)

3.1 regulations and regulators (includes price caps)
3.2 internal projects -- progress and strategy
3.3 company image -- current
3.4 company image -- changing / influencing
3.5 ...

4 Emotional tone (if not neutral) (19)

4.1 jubilation
4.2 hope / anticipation
4.3 humor
4.4 camaraderie
4.5 admiration
4.6 gratitude
4.7 friendship / affection
4.8 sympathy / support
4.9 sarcasm
4.10 secrecy / confidentiality
4.11 ...

Data categories

The original UCB System had to be broken down for this pilot, but more categories can be technically be added.

1 Coarse genre

1.1 Company Business, Strategy, etc. ~~(elaborate in Section 3 [Topics])~~
1.2 Purely Personal
1.3 Personal but in professional context (e.g., it was good working with you)
1.4 Logistic Arrangements (meeting scheduling, technical support, etc)
1.5 Employment arrangements (job seeking, hiring, recommendations, etc)
1.6 Document editing/checking (collaboration)
~~1.7 Empty message (due to missing attachment)~~
~~1.8 Empty message~~
1.7 Payroll, finance, accounting etc.

Data pre-processing

Message-ID: <32084772.1075857584968.JavaMail.evans@thyme>Date: Sun, 10 Dec 2000 22:05:00 -0800 (PST)From: jennifer.fraser@enron.comTo: russell.dyk@enron.comSubject: LNG QuestionsCc: john.arnold@enron.comMime-Version: 1.0Content-Type: text/plain; charset=us-asciiContent-Transfer-Encoding: 7bitBcc: john.arnold@enron.comX-From: Jennifer FraserX-To: Russell DykX-cc: John ArnoldX-bcc: X-Folder: \John_Arnold_Dec2000\Notes Folders\Notes inboxX-Origin: Arnold-JX-FileName: Jarnold.nsfRuss:A couple of questions1- Check the DOE Northeast Heating season report. There seem to be a lot of LNG terminal and facilities in the Northeast. How do they work? What are the logistics of transportation etc...2- Arnold's buddy has been looking into the logistics of trucking LNG to CA. Is this possible? Can investigate the probability?ThanksJF

Before

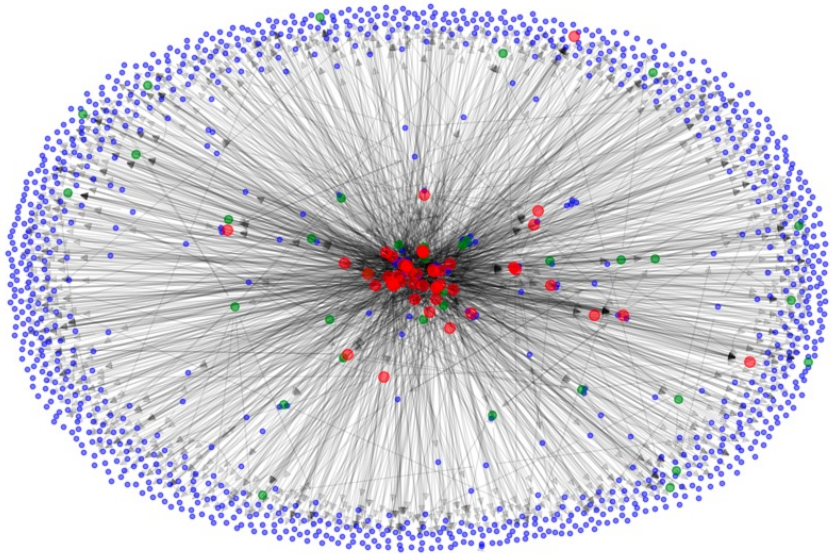
Regular Expressions (RE) + Stop Words Removal + Lemmatization + Stemming

rus couple question check doe northeast heating season report seem lot lng terminal facility northeast do they work logistics transportation Arnold buddy been looking into logistics trucking lng ca possible can investigate probability thanks jf

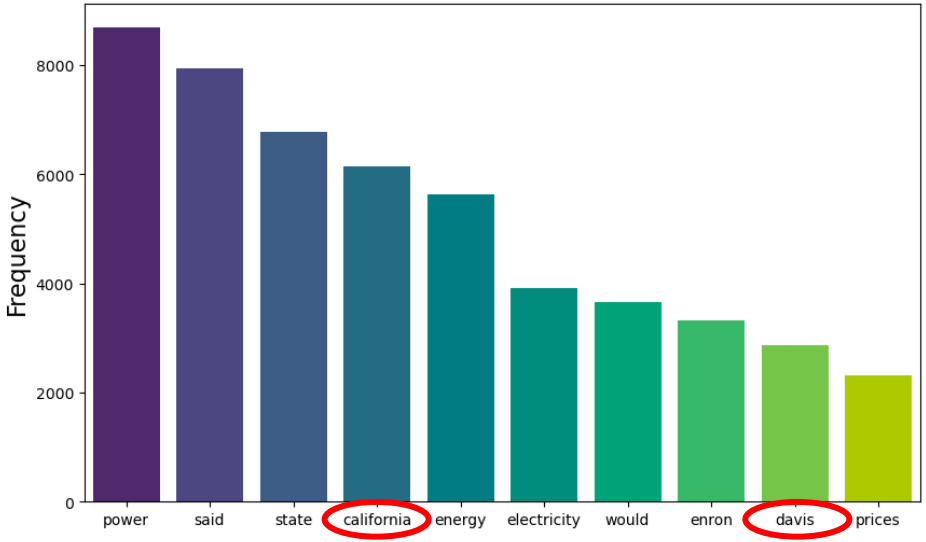
After

Data visualization

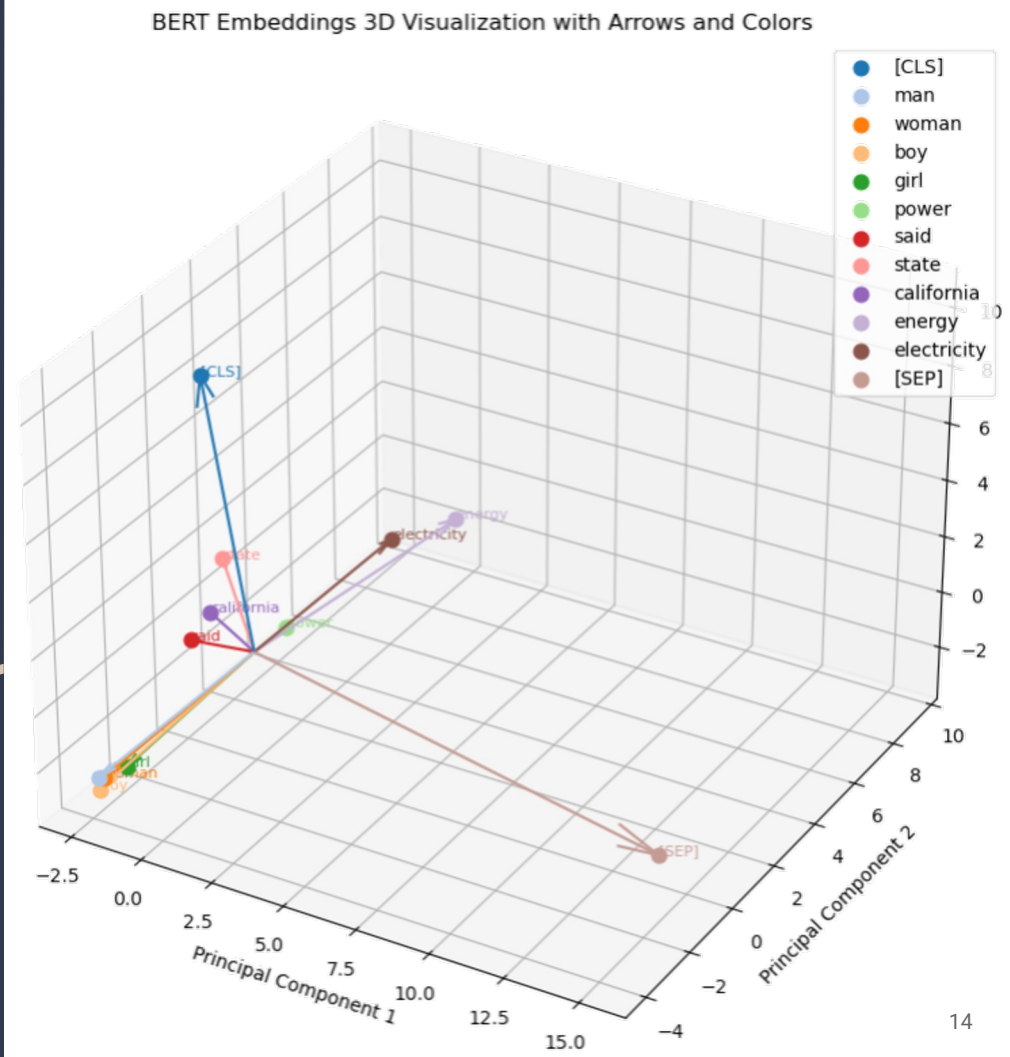
Email Network (UCB Dataset)



Frequency of contained words (UCB Dataset)



3. The findings



Model selection

Llama2-7b

- Introduced by Meta
- 7 billion parameters
- State of the art - released in 2023
- Open source – good community support
- Decoder only architecture – text generation

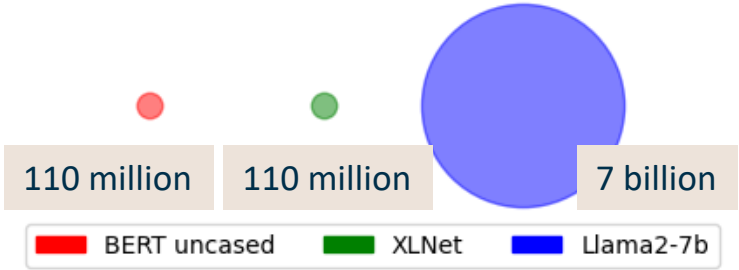
XLNet base

- Introduced by Carnegie Mellon University
- 110 million parameters
- Released in 2019
- Design to improve upon BERT uncased
- Bidirectional context

BERT base uncased

- Introduced by Google
- 110 million parameters
- Released in 2018
- Open source – good community support
- Bidirectional encoder architecture
- Either left-to-right or right-to-left context

Parameter comparison



Model evaluation

Testing always against UCB labels	**LLama2 Fine-Tuned**	**BERT Fine-Tuned***	**XLNet Fine-Tuned***	**LLama2 general**	**ChatGPT 3.5 general**	**Manual**
Acc. on test set	**65%** *n~70*	**67%** *n~3350*	**62%** *n~3350*	**43%** *n=198*	**53%** *n=75*	**55%** *n=200*
Training time†	~2.5 h	~4 h	~6 h	/	/	/
Time for one answer†	7 sec	<1 sec	<1 sec	9 sec	<1 sec	24 sec

Addendum

- One email often fits into several categories, even when just considering “Coarse Genre”, making a measurement of accuracy particularly challenging.
- Fine-Tuning can bring benefits to current state-of-art architectures but is computationally expansive.
- The size of the training dataset drastically influences performance.

*These models were trained and tested with additional ChatGPT 3.5 labeled data.
†Tested with very limited computing power, can easily be substantially improved.

Summary and future outlook

Key findings

1. Manual annotation is expensive and can be replaced by AI methods at lower cost and comparable accuracy.
2. Categorization accuracy is hard to measure, as several categories are often applicable to one email (At least for the Enron example).
3. A lot of data cleaning can be required if the structure of emails changes in any way.
4. Fine-Tuning models is computationally expensive, but improvements are possible even compared to state-of-art LLM-models.

Next steps

1. Incoming emails should be stored in a consistent format, to avoid high costs for data cleaning.
2. More categories should be tested and included than in this pilot.
3. Bigger models (>7b parameters) will outperform the accuracies achieved by us.
4. Larger training datasets for fine-tuning and computational resources will contribute towards higher accuracies.
5. Application on the real DB email dataset.

Key references

- Enron dataset labelling project (UC Berkeley): https://bailando.berkeley.edu/enron_email.html
- Goh, Y. C., Cai, X. Q., Theseira, W., Ko, G., & Khor, K. A. (2020). Evaluating human versus machine learning performance in classifying research abstracts. Scientometrics, 125(2), 1197–1212. https://doi.org/10.1007/s11192-020-03614-2
- Labelled Enron dataset: https://data.world/brianray/enron-email-dataset
- Network depiction of complete Enron dataset: https://cambridge-intelligence.com/using-social-network-analysis-measures/
- Parent Llama LLM: https://huggingface.co/meta-llama/Llama-2-7b-hf
- Parent Bert Base Uncased LLM: https://huggingface.co/bert-base-uncased
- Parent XLNet Base LLM: https://huggingface.co/docs/transformers/model_doc/xlnet

Further resources

https://github.com/neelblabla/large_language_models_for_processing_emails

https://huggingface.co/neelblabla/email-classification-llama2-7b-peft

Appendix

	message_id	subject	content
0	<6184043.1075857585289.JavaMail.evans@thyme>	more about you	A little bird told me about this. Read it wh...
1	<17858209.1075857584947.JavaMail.evans@thyme>	LNG on the road	trucking LNG to California - The short ans...
2	<14187877.1075857584924.JavaMail.evans@thyme>	Amazon.com Password Assistance	Greetings from Amazon.com. To finish resetti...
3	<31194989.1075857585233.JavaMail.evans@thyme>	Deferral Enrollment 2001	...
4	<32084772.1075857584968.JavaMail.evans@thyme>	LNG Questions	Russ: A couple of questions - Check the DOE ...
...	...	...	...
481235	<13299843.1075840028954.JavaMail.evans@thyme>	RE: WSCC Tagging practices	I agree with Bob on #. I am indifferent on # ...
481236	<13266154.1075840029143.JavaMail.evans@thyme>	E-Tag Version 1.7 Implementation Delayed	TO: CONTROL AREAS TRANSMISSION PROVIDER...
481237	<13610289.1075840029369.JavaMail.evans@thyme>	RE: Common Scheduling Time Zone across WSCC	Good Morning Diana I just spoke with Duong a...
481238	<21704474.1075840029683.JavaMail.evans@thyme>	Late tags	Nothing is easy is it?? At the risk of the c...
481239	<9367927.1075840029633.JavaMail.evans@thyme>	RE: BCHA Automatic Denial/Approval	I think you are right on! In addition I would...

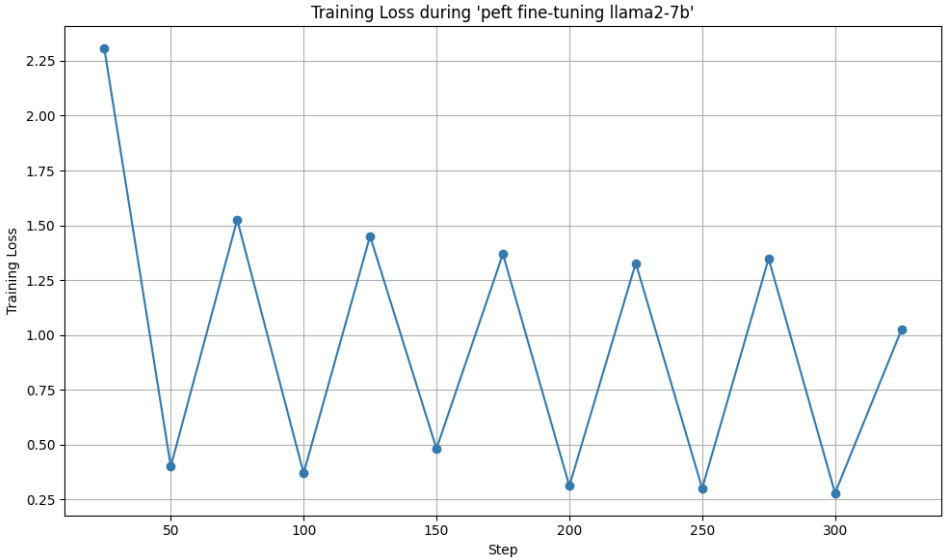
Appendix

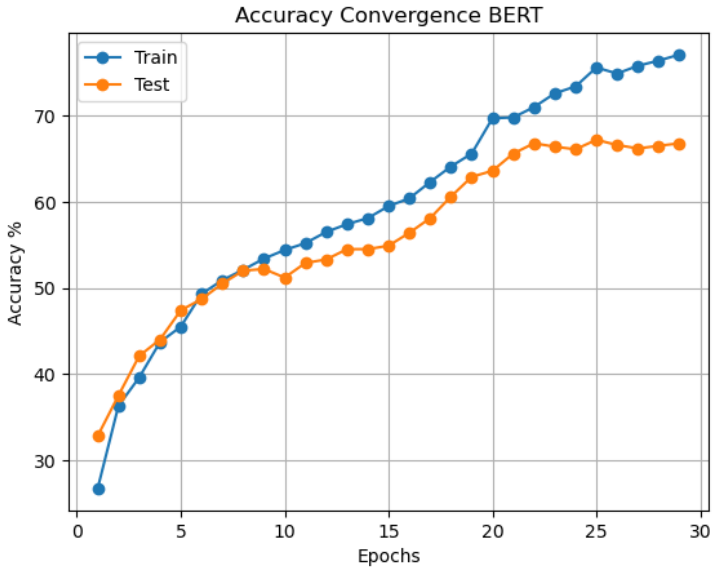
Model architectures

Parameters	BERT base 110m (fine-tuned)	XLNet base 110m (fine-tuned)	Llama2 7b (base)
Framework	PyTorch	PyTorch	PyTorch
Total Attention Layers	12	12	32
Frozen Layers	8 (first 8)	8 (first 8)	/
Active Layers	4 (last 4)	4 (last 4)	/
Batch Size	30	38	4M
Train / Test split	75 / 25	75 / 25	/
Norm	L1 norm	-	RMSNorm
Learning Rate	0.000005	0.00001	0.00001
Epochs	~50	~50	~1T tokens
Optimizer	Adam	Adam	AdamW
Loss	Cross Entropy	Cross Entropy	Cosine Loss

Appendix

Model performances*





*Generative models (Llama2) are evaluated using loss, while categorical models (BERT) use accuracy.

Appendix

Prompting Llama

Training Prompts = Instruction + {Mail} + {Category}

```
"<s>[INST] <<SYS>> I am sharing an email body with you. Based on the text in the body, you need to classify the email in one of the following eight categories: 'Company Business, Strategy, etc.'; 'Purely Personal'; 'Personal but in professional context (e.g., it was good working with you)'; 'Logistic Arrangements (meeting scheduling, technical support, etc)'; 'Employment arrangements (job seeking, hiring, recommendations, etc)'; 'Document editing/checking (collaboration)'; 'Empty message (due to missing attachment)'; 'Empty message'. <</SYS>> Mail: Aruna I shall be in London this week. Please call me on Monday next week. Best time is between 7:30 and 8:30 my time. Vince [/INST] Category: Company Business, Strategy, etc.(elaborate in Section 3 [Topics])"
```

Testing Prompts = Instruction + {Mail}

```
"<s>[INST] <<SYS>> I am sharing an email body with you. Based on the text in the body, you need to classify the email in one of the following eight categories: 'Company Business, Strategy, etc.'; 'Purely Personal'; 'Personal but in professional context (e.g., it was good working with you)'; 'Logistic Arrangements (meeting scheduling, technical support, etc)'; 'Employment arrangements (job seeking, hiring, recommendations, etc)'; 'Document editing/checking (collaboration)'; 'Empty message (due to missing attachment)'; 'Empty message'. <</SYS>> Mail: Put FH on AAE retainer Bd Meeting on Sun in Bd Rm at 8:00am Mary Kay -- posi'n on TW: ROW issues call Leslie [/INST] Category: "
```

Appendix

Model fine-tuning