Investigating False Negatives for LLMs

Goal: Current LLM models have non-zero FNR, but we need them due to their high TNRs.

Task: Lower FNR to zero

Methods

We investigate the false negatives in Sonnet zero-shot. There is only one false negative. Here is the copy:

Email metadata

- Email ID: 109
- Subject: TRANSPORTATION MODEL
- From: Phillip K Allen
- To: Keith Holst
- Ground truth PII: full name

Full message body

```
---------------------- Forwarded by Phillip K Allen/HOU/ECT on 08/09/2000
02:11 PM ---------------------------

   Enron North America Corp.

   From:  Colleen Sullivan                    08/09/2000 10:11 AM


To: Keith Holst/HOU/ECT@ect, Andrew H Lewis/HOU/ECT@ECT, Fletcher J
Sturm/HOU/ECT@ECT, Larry May/Corp/Enron@Enron, Kate Fraser/HOU/ECT
@ECT, Zimin
```

```
Lu/HOU/ECT@ECT, Greg Couch/HOU/ECT@ECT, John Griffith/Corp/Enron@E
nron,
Sandra F Brawner/HOU/ECT@ECT, John J Lavorato/Corp/Enron@Enron, Hunt
er S
Shively/HOU/ECT@ECT, Phillip K Allen/HOU/ECT@ECT, Scott Neal/HOU/ECT
@ECT,
Thomas A Martin/HOU/ECT@ECT, Steve Jackson/HOU/ECT@ECT
cc: Julie A Gomez/HOU/ECT@ECT, Stephanie Miller/Corp/Enron@ENRON
Subject: TRANSPORTATION MODEL

Please plan to attend a meeting on Friday, August 11 at 11:15 a.m. in 30C1 to
discuss the transportation model.  Now that we have had several traders
managing transportation positions for several months, I would like to discuss
any issues you have with the way the model works.   I have asked Zimin Lu
(Research), Mark Breese and John Griffith (Structuring) to attend so they
will be available to answer any technical questions.   The point of this
meeting is to get all issues out in the open and make sure everyone is
comfortable with using the model and position manager, and to make sure tho
se
who are managing the books believe in the model's results.   Since I have
heard a few concerns, I hope you will take advantage of this opportunity to
discuss them.

Please let me know if you are unable to attend.
```

Here is a recap on the PII list we care about

```
PII Element,Notes
[full name],"Only flag [full name] if it appears in combination with another PII e
lement, NOT inclusive of [email address]."
[ssn],Social security number
"[ssn, last 4]",Social security number -- last 4 digits
[drivers license #],
[passport #],
[TIN],Taxpayer identification number
[irs identity protection pin],
[student identification #],
[bank account #],"Account number, NOT routing number"
[credit card #],Includes credit and debit cards
[cvv/cvc],Credit or debit card security pin
[password],
[username],
[email address],
[phone number],
```

We note that the ground truth label is `full name`, but in the PII list, we will only care about full name if it occurs in combination with other non eamail address PII. By manual inspection, this email doesn’t have that combination, and I consider this to be a incorrect golden label.

For Haiku, we note 3 FN emails. One of it is above. The other two are below:

False Negative #1: Email ID 7

Email Metadata:

- **Subject**: (empty)
- **From**: Phillip K Allen
- **To**: david.l.johnson@enron.com, John Shafer
- **Ground Truth PII**: email address

Full Message Body:

```
Please cc the following distribution list with updates:

Phillip Allen (pallen@enron.com)
Mike Grigsby (mike.grigsby@enron.com)
Keith Holst (kholst@enron.com)
Monique Sanchez
```

```
Frank Ermis
John Lavorato

Thank you for your help

Phillip Allen
```

False Negative #3: Email ID 242

Email Metadata:

- **Subject**: PIRA's California/Southwest Gas Pipeline Study
- **From**: Phillip K Allen
- **To**: Mike Grigsby
- **Ground Truth PII**: full name, phone number

Full Message Body:

```
---------------------- Forwarded by Phillip K Allen/HOU/ECT on 02/21/2000=
=20
08:06 AM ---------------------------
  =20
=09
=09
=09From:  Jennifer Fraser                     02/19/2000 01:57 PM
=09

To: Stephanie Miller/Corp/Enron@ENRON, Julie A Gomez/HOU/ECT@ECT, Phillip K=
=20
Allen/HOU/ECT@ECT
cc: =20
Subject: PIRA's California/Southwest Gas Pipeline Study

Did any of you order this
```

```
JEn

--------------------- Forwarded by Jennifer Fraser/HOU/ECT on 02/19/2000
=
=20
03:56 PM ---------------------------


"Jeff Steele" <jsteele@pira.com> on 02/14/2000 01:51:00 PM
To: "PIRA Energy Retainer Client" <jsteele@pira.com>
cc: =20
Subject: PIRA's California/Southwest Gas Pipeline Study

...

If you have any questions, please do not hesitate to  contact me.

Sincerely,

Jeff Steele
Manager, Business  Development
PIRA Energy Group
(212) 686-6808
jsteele@pira.com =20

 - PROSPECTUS.PDF
 - PROSPECTUS.doc
```

This shows that Haiku is indeed unreliable.

Takeaways

We can treat Sonnet zero shot as the ground truth. Haiku is unreliable. An idea from this analysis is that we can assist the LLM by feeding in the entities flagged by Presidio, and also ask the model to output confidence score from 1 to 5, 5 being very sure it is a PII of concern.