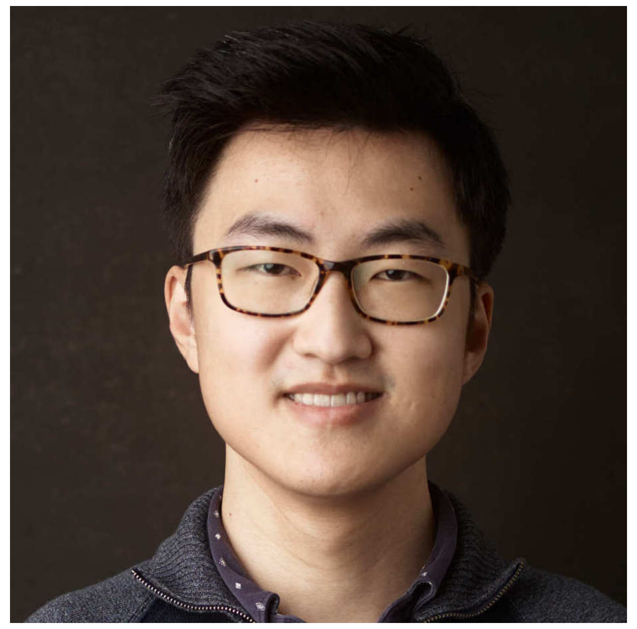
Standardizing and unifying unstructured instrument data with LLMs

Andrew Chen



Andrew Chen

Stanford Y Combinator

Stanford MS Computer Science at age 20

Florence Pham

airbnb YouTube

Tech lead for highest-traffic page at Airbnb
>100M visitors

The Problem

Analogy: Translation

English

Analogy: Translation

Mandarin

English

Analogy: Translation

Mandarin

Catalan

English

Analogy: Translation

Mandarin

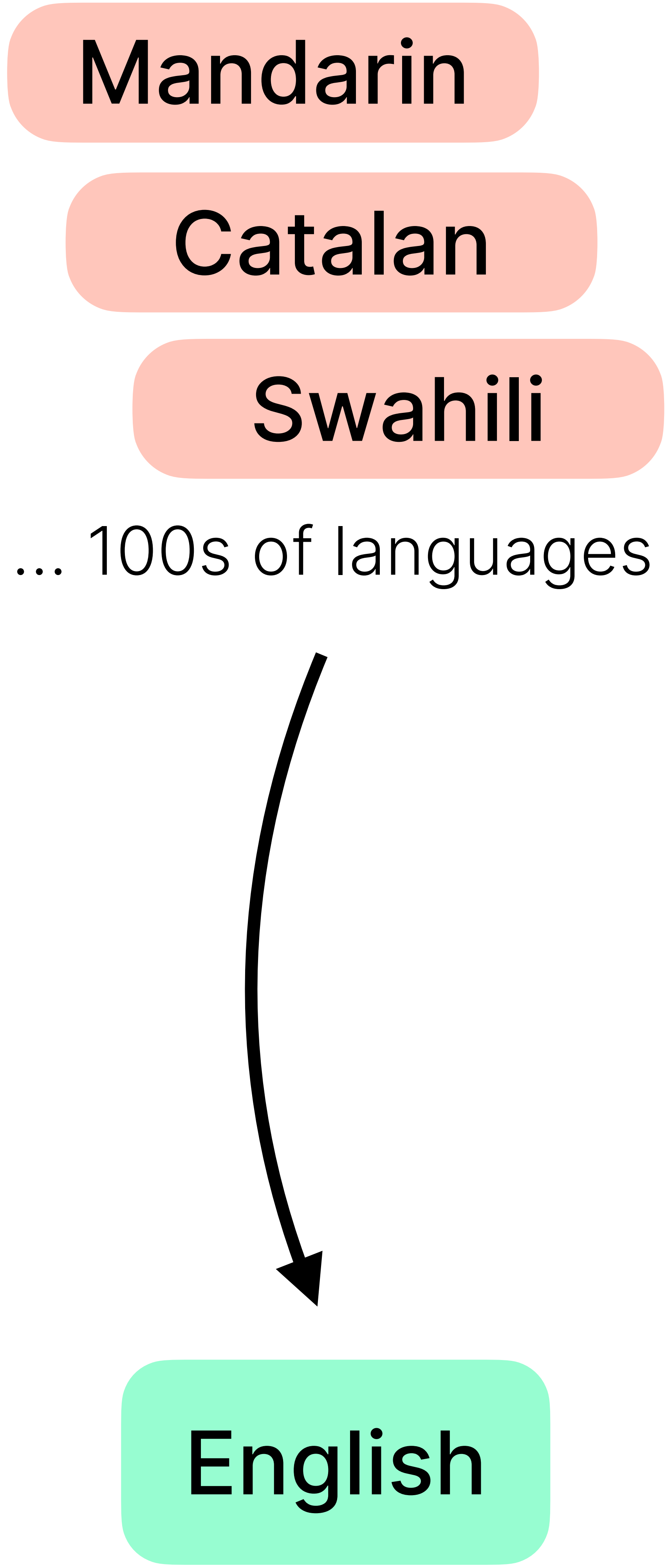
Catalan

Swahili

... 100s of languages

English

Analogy: Translation



Mandarin

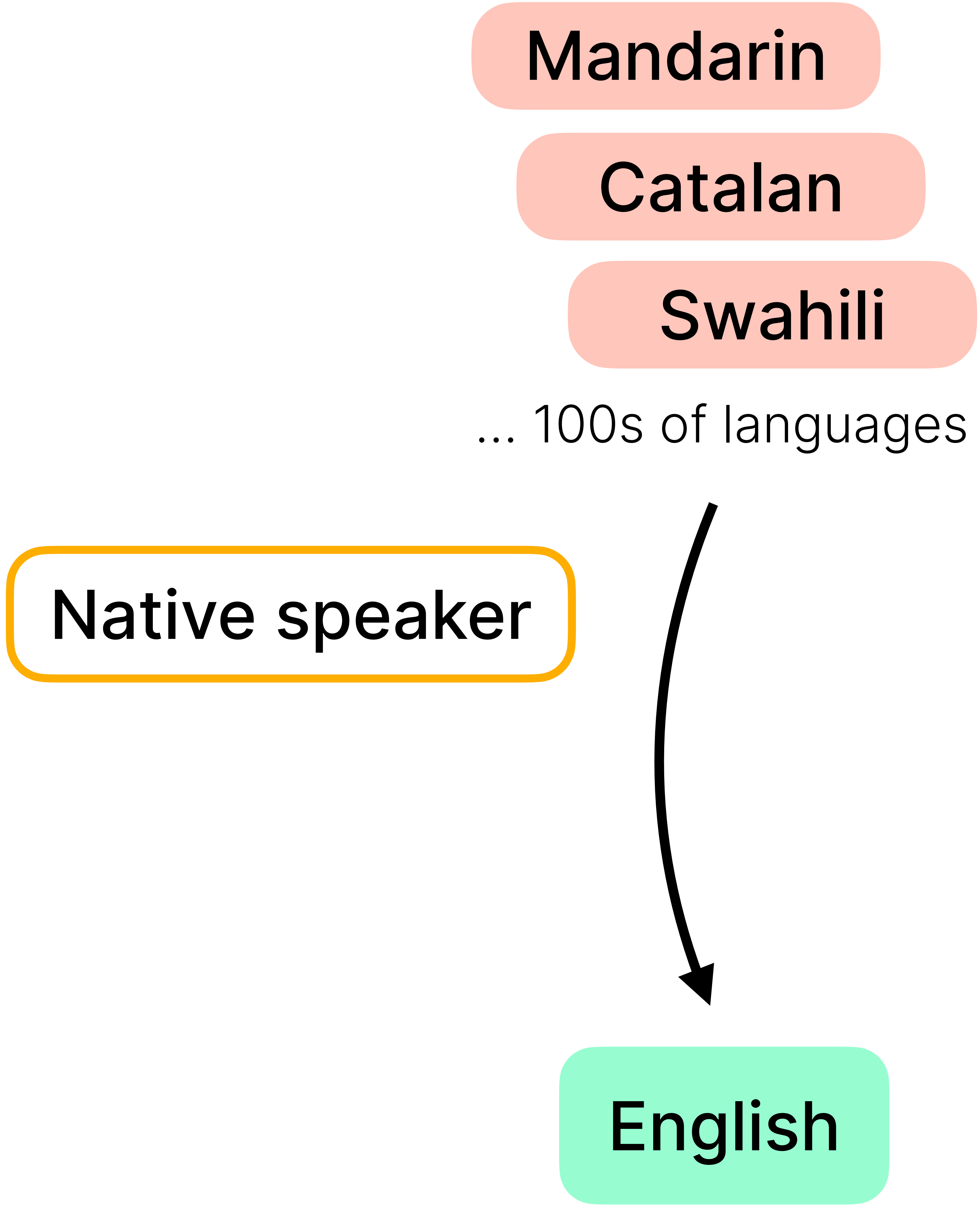
Catalan

Swahili

... 100s of languages

English

Analogy: Translation



Mandarin

Catalan

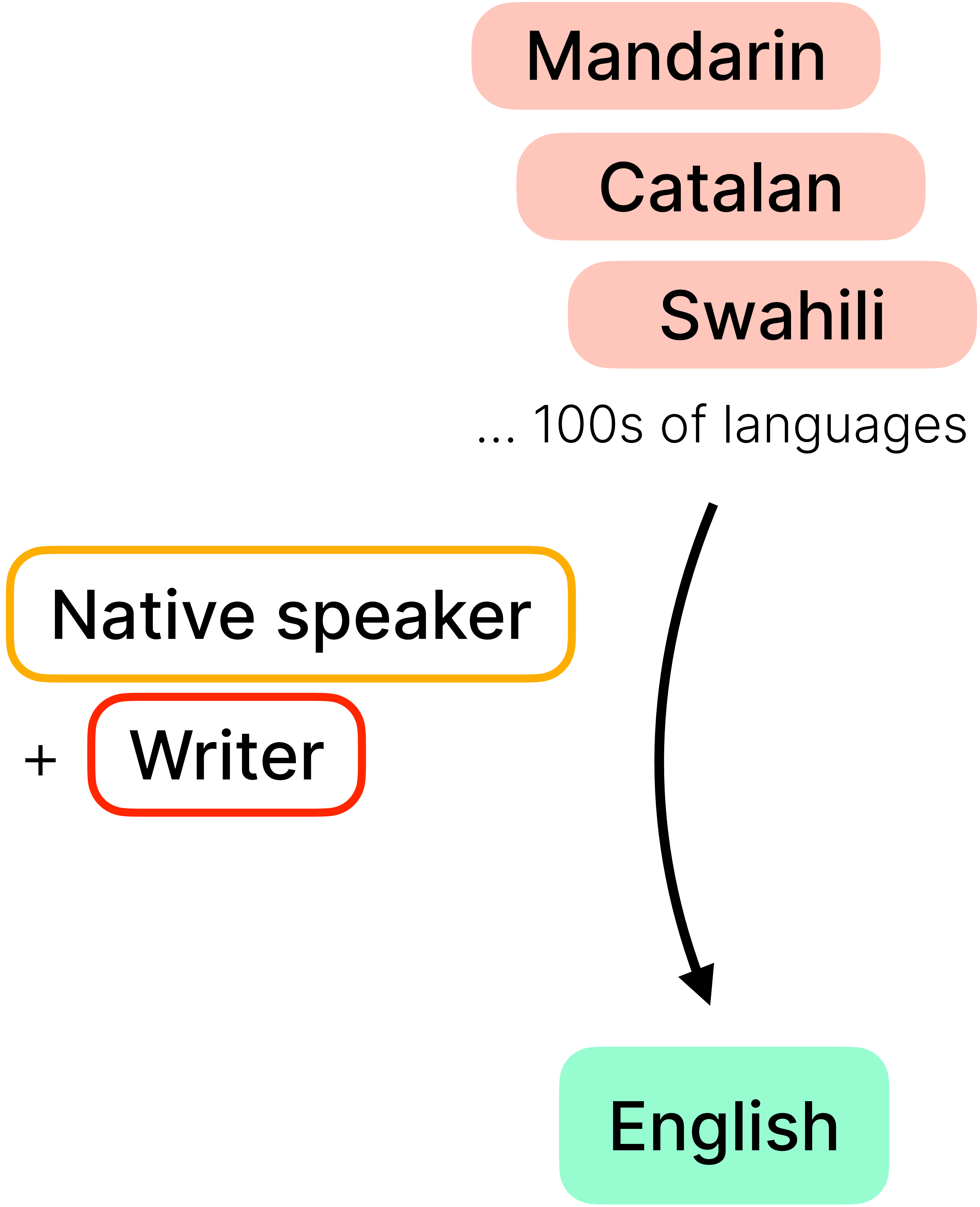
Swahili

... 100s of languages

Native speaker

English

Analogy: Translation



Mandarin

Catalan

Swahili

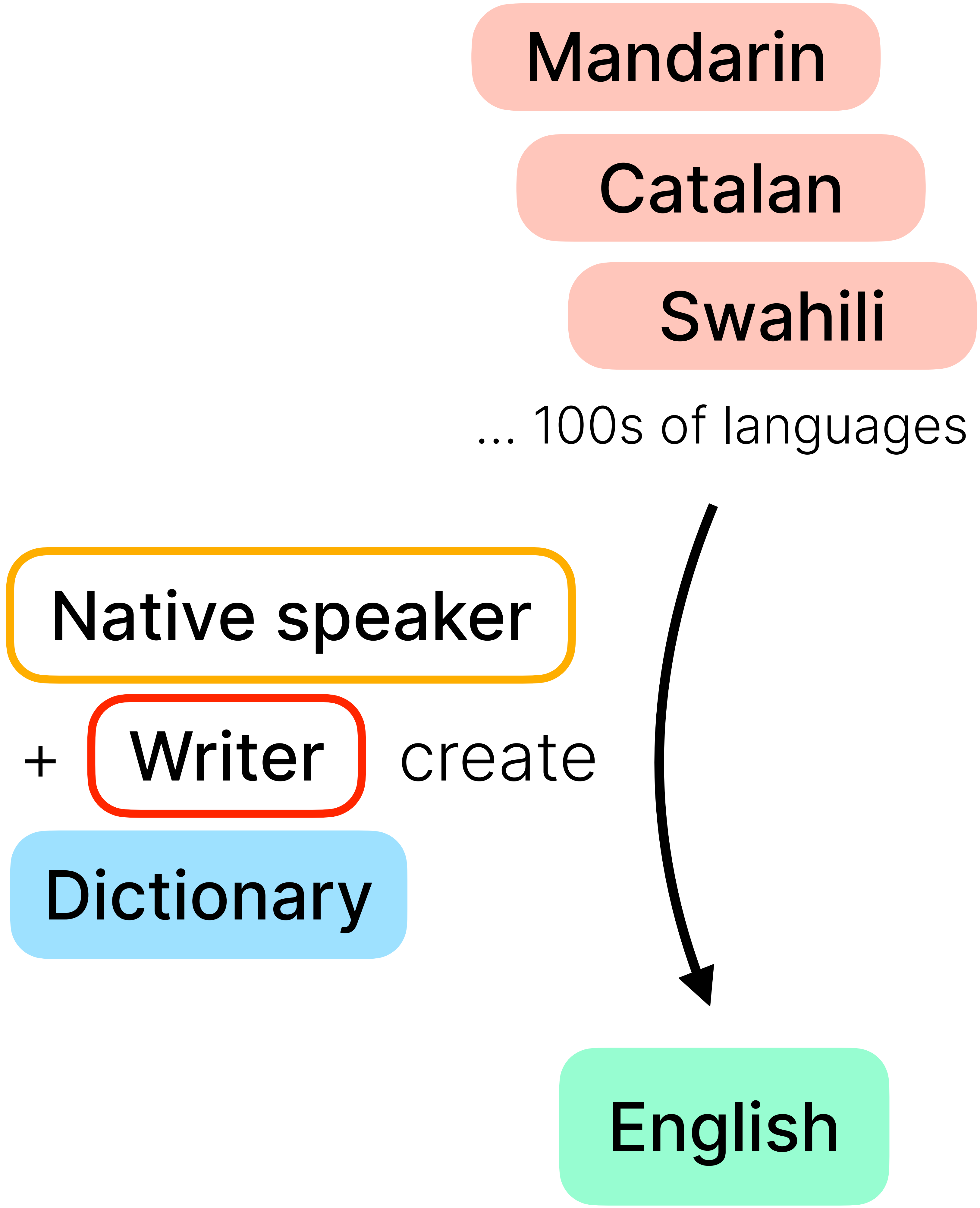
... 100s of languages

Native speaker

+ Writer

English

Analogy: Translation



Mandarin

Catalan

Swahili

... 100s of languages

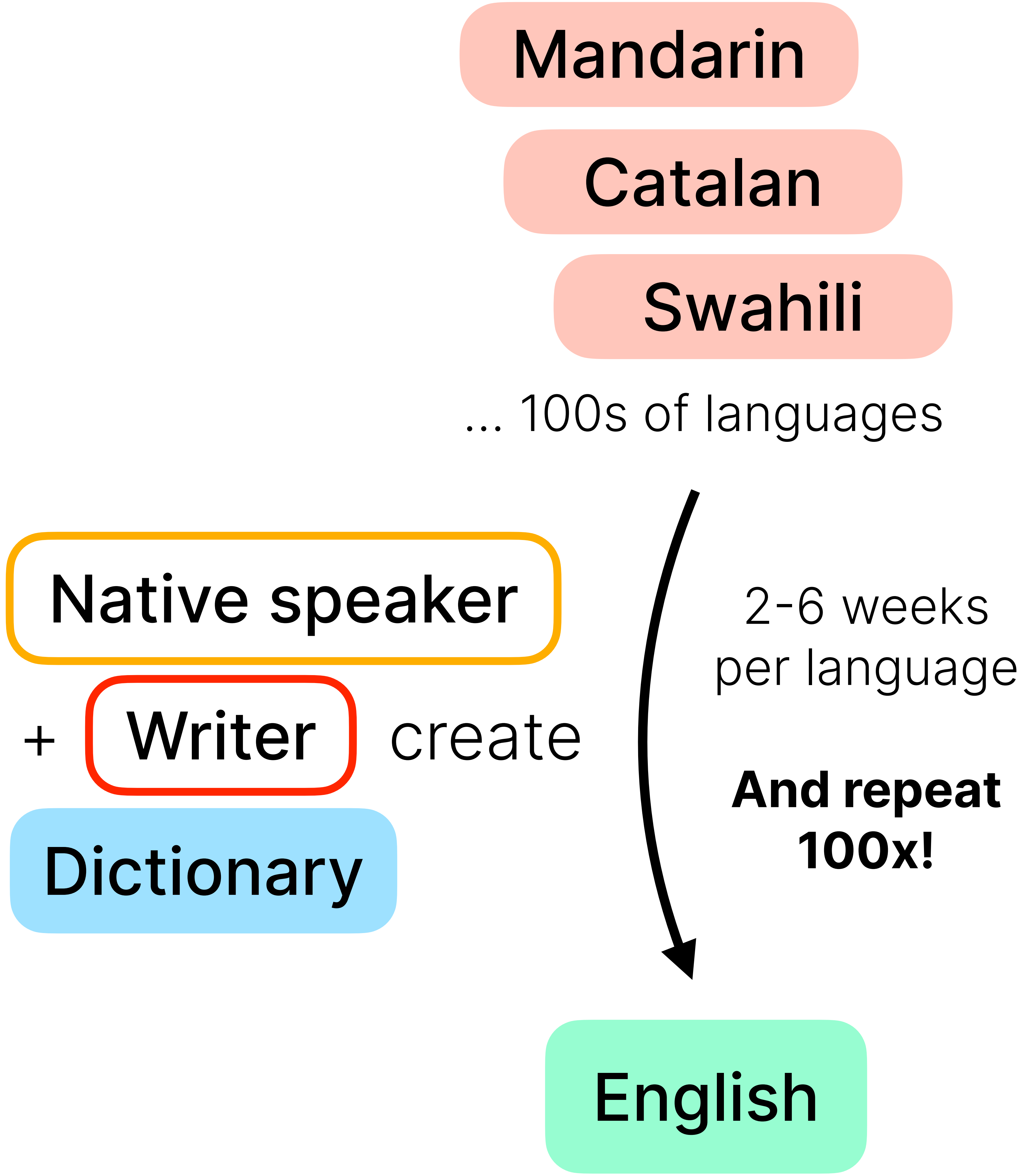
Native speaker

+ Writer create

Dictionary

English

Analogy: Translation



Mandarin

Catalan

Swahili

... 100s of languages

Native speaker

+ Writer create

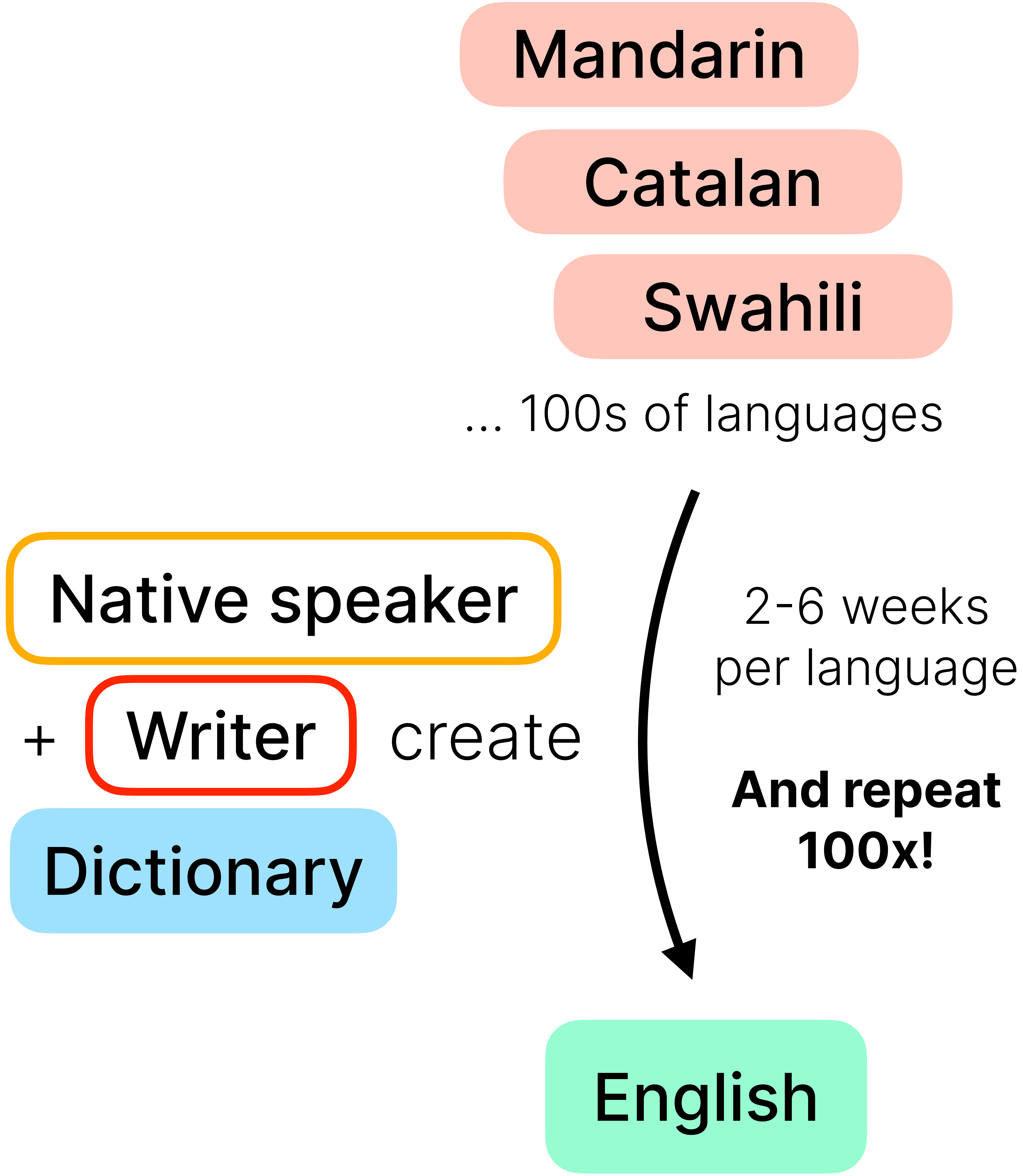
Dictionary

2-6 weeks
per language

**And repeat
100x!**

English

Analogy: Translation



Mandarin

Catalan

Swahili

... 100s of languages

Native speaker

+ Writer create

Dictionary

2-6 weeks per language

And repeat 100x!

English

Status Quo

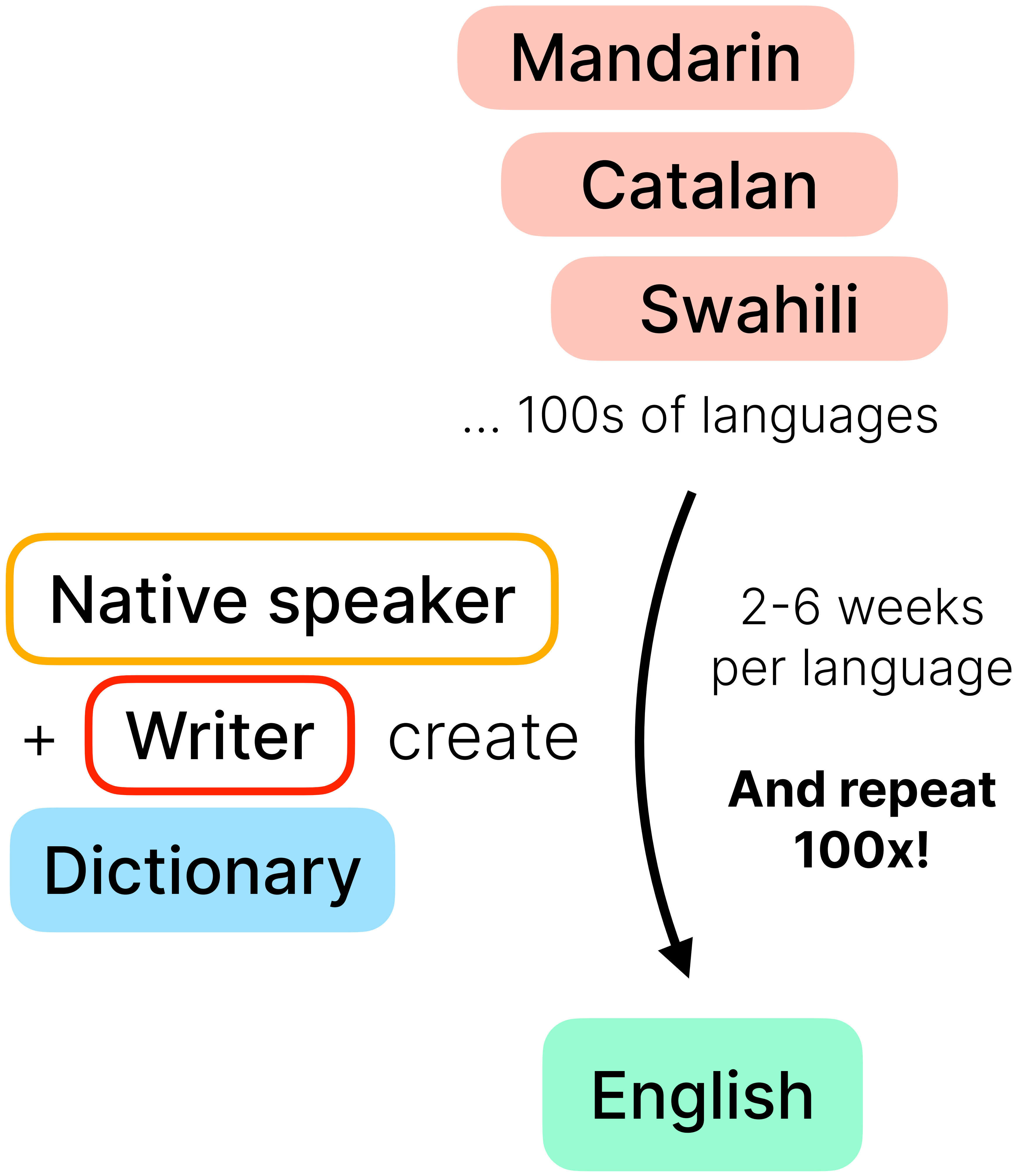
Proprietary

Instrument

Formats

... 1000s of formats

Analogy: Translation



Mandarin

Catalan

Swahili

... 100s of languages

Native speaker

+ Writer create

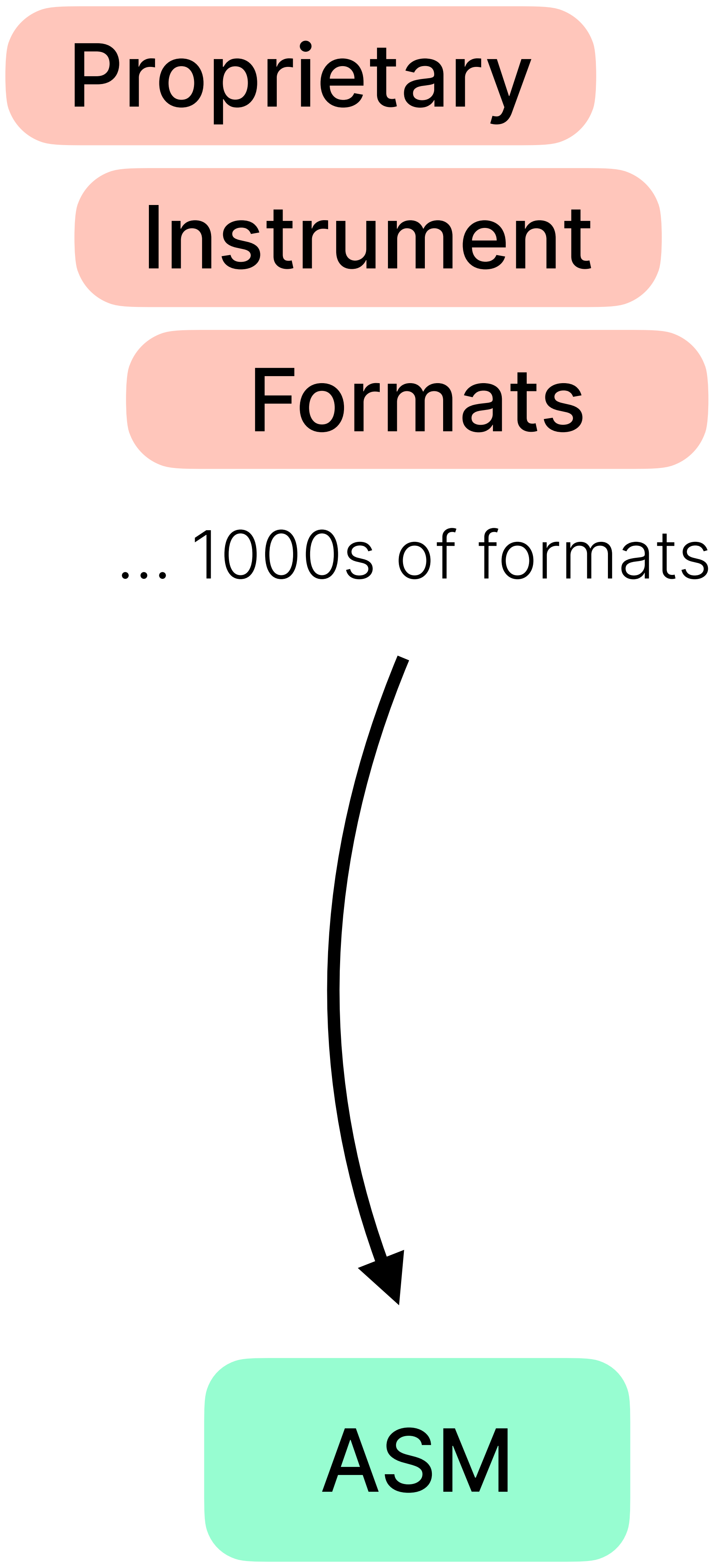
Dictionary

2-6 weeks
per language

**And repeat
100x!**

English

Status Quo



Proprietary

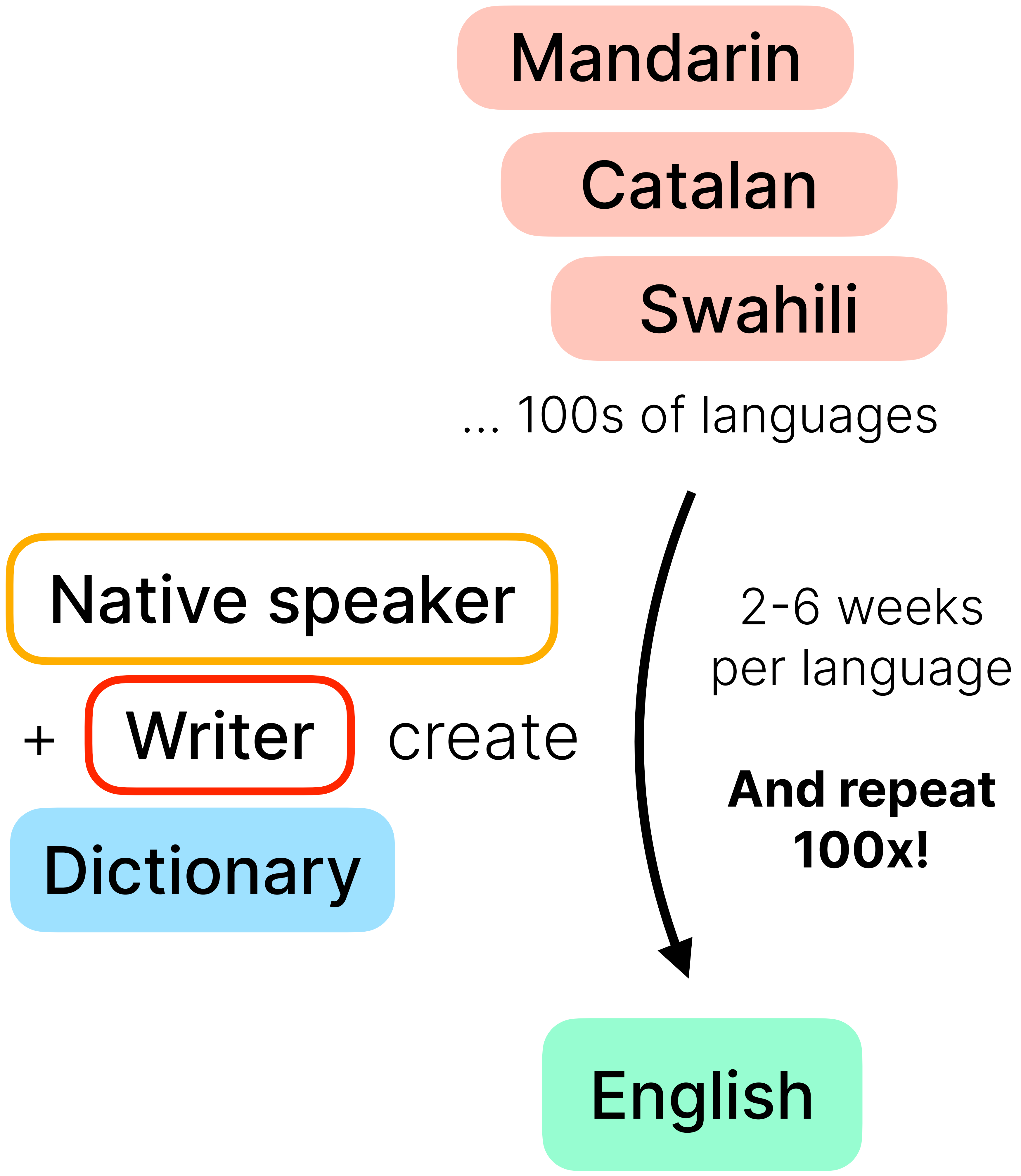
Instrument

Formats

... 1000s of formats

ASM

Analogy: Translation



Mandarin

Catalan

Swahili

... 100s of languages

Native speaker

+ Writer create

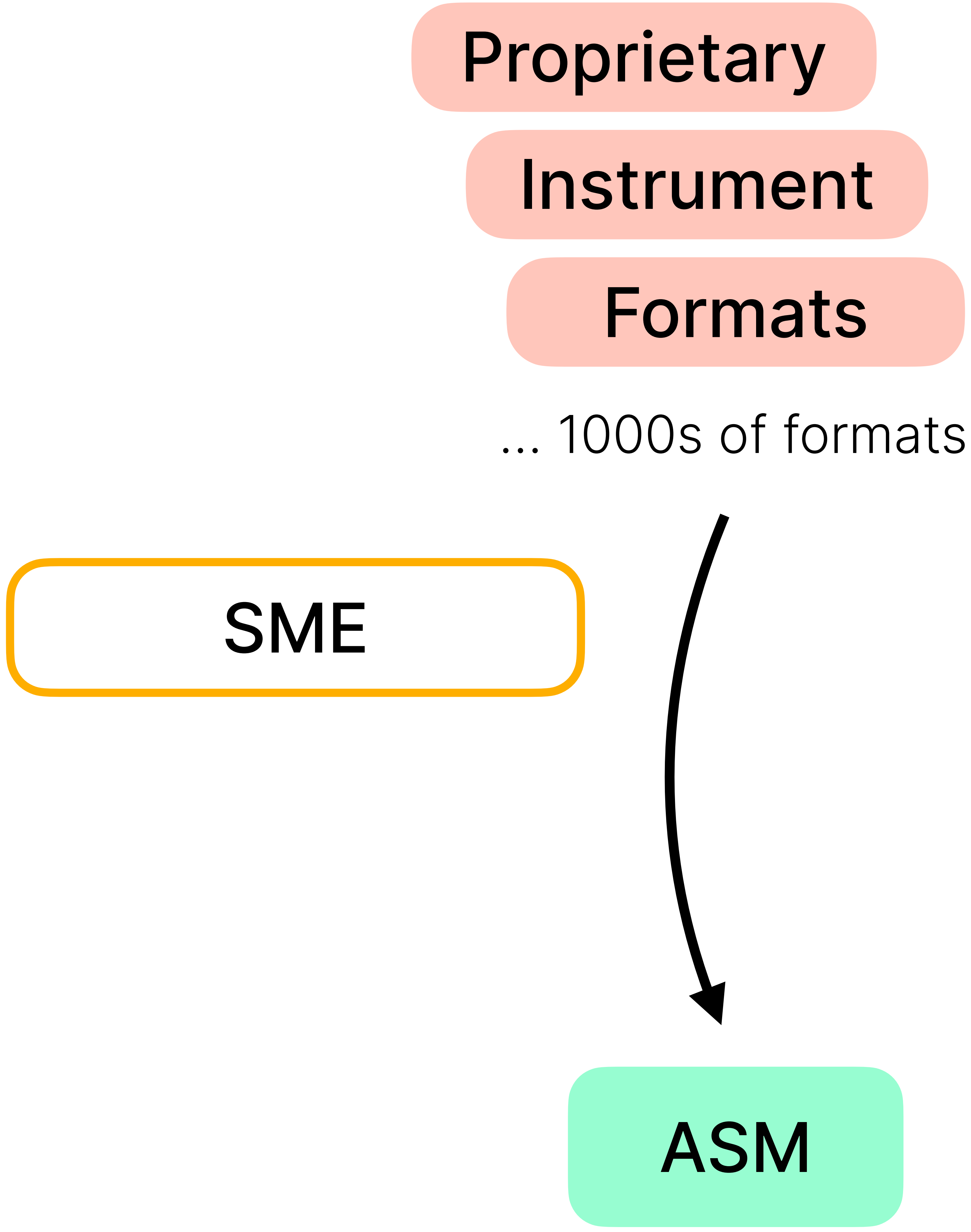
Dictionary

2-6 weeks
per language

And repeat
100x!

English

Status Quo



Proprietary

Instrument

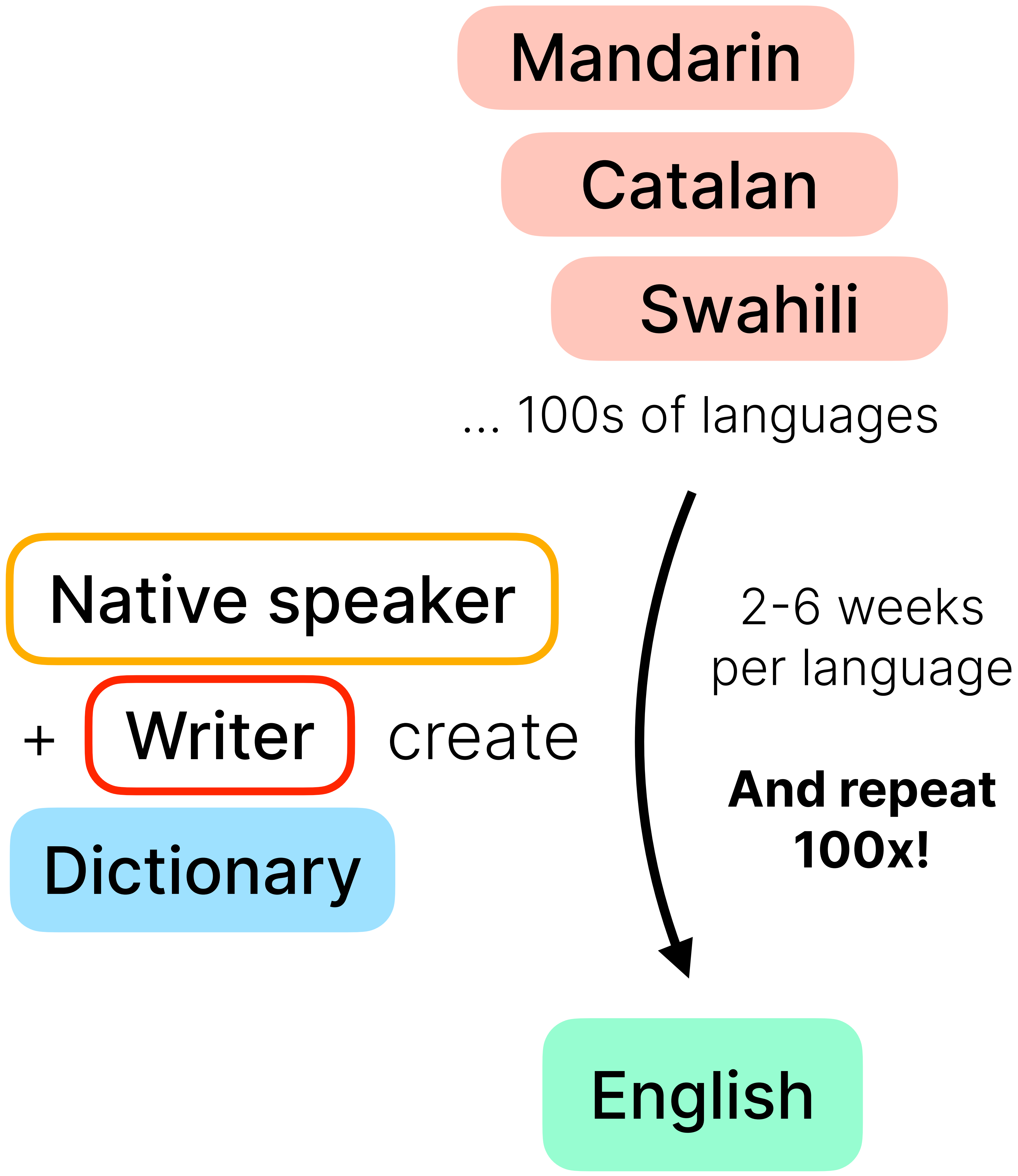
Formats

... 1000s of formats

SME

ASM

Analogy: Translation



Mandarin

Catalan

Swahili

... 100s of languages

Native speaker

+ Writer create

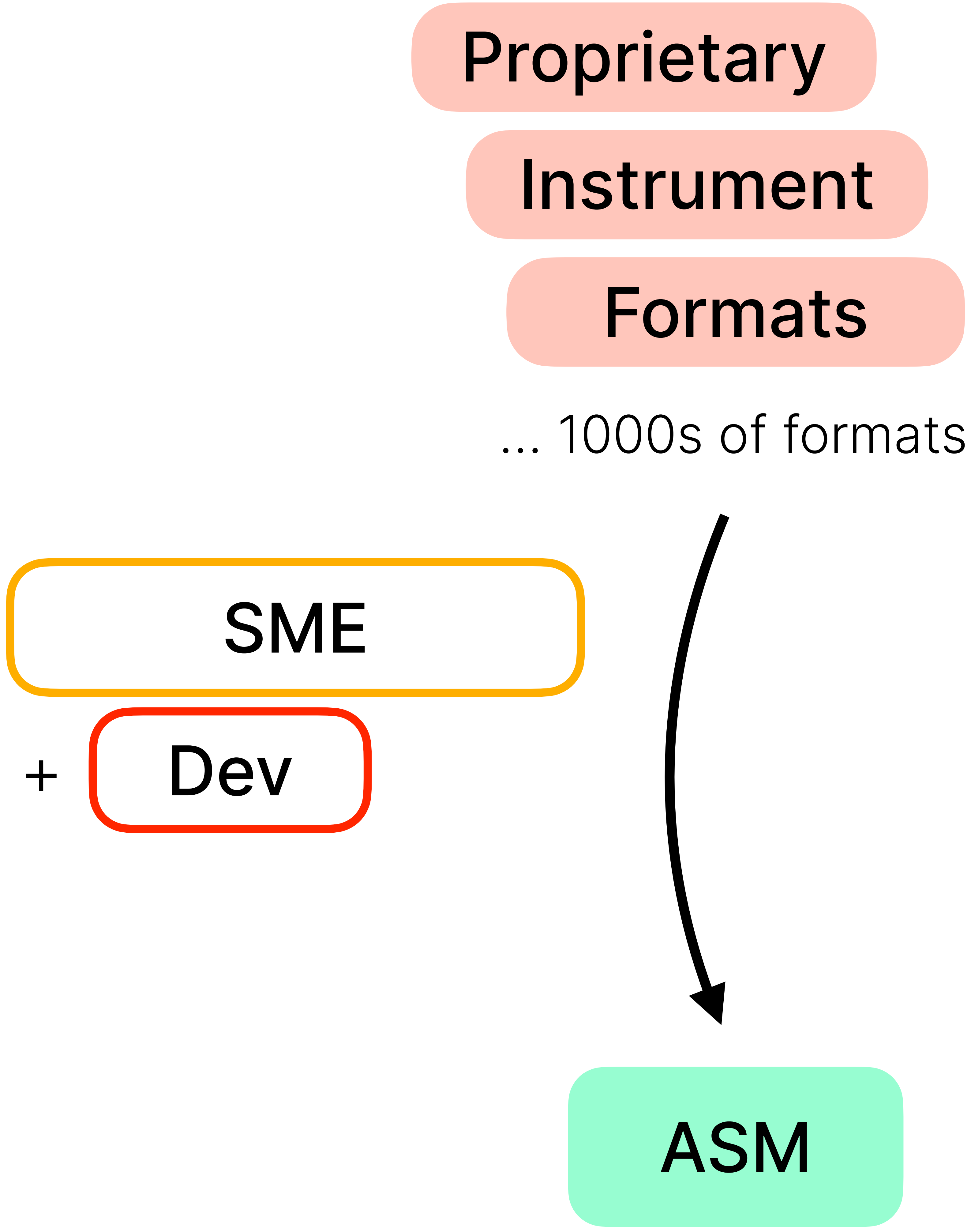
Dictionary

2-6 weeks per language

And repeat 100x!

English

Status Quo



Proprietary

Instrument

Formats

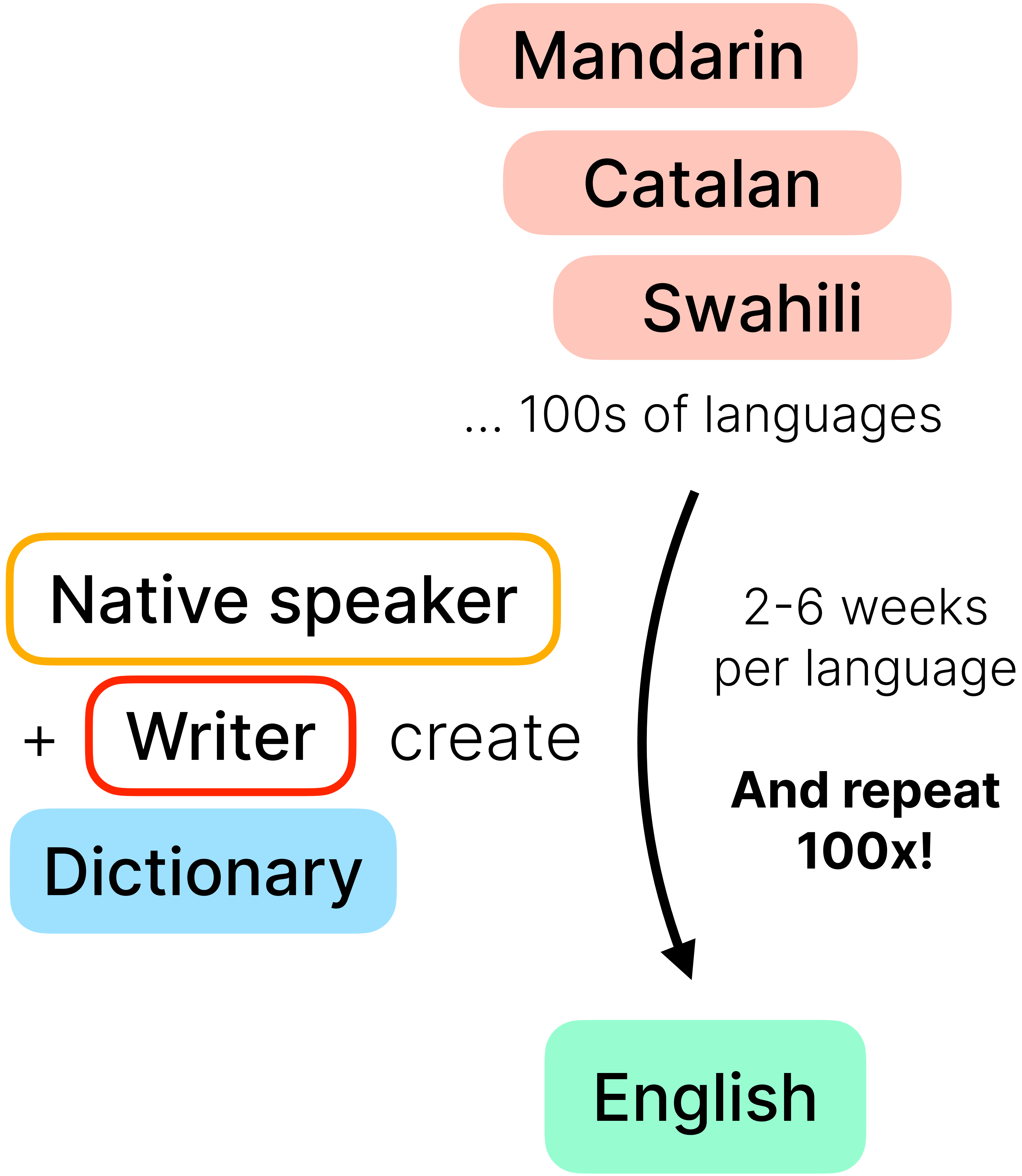
... 1000s of formats

SME

+ Dev

ASM

Analogy: Translation



Mandarin

Catalan

Swahili

... 100s of languages

Native speaker

+ Writer create

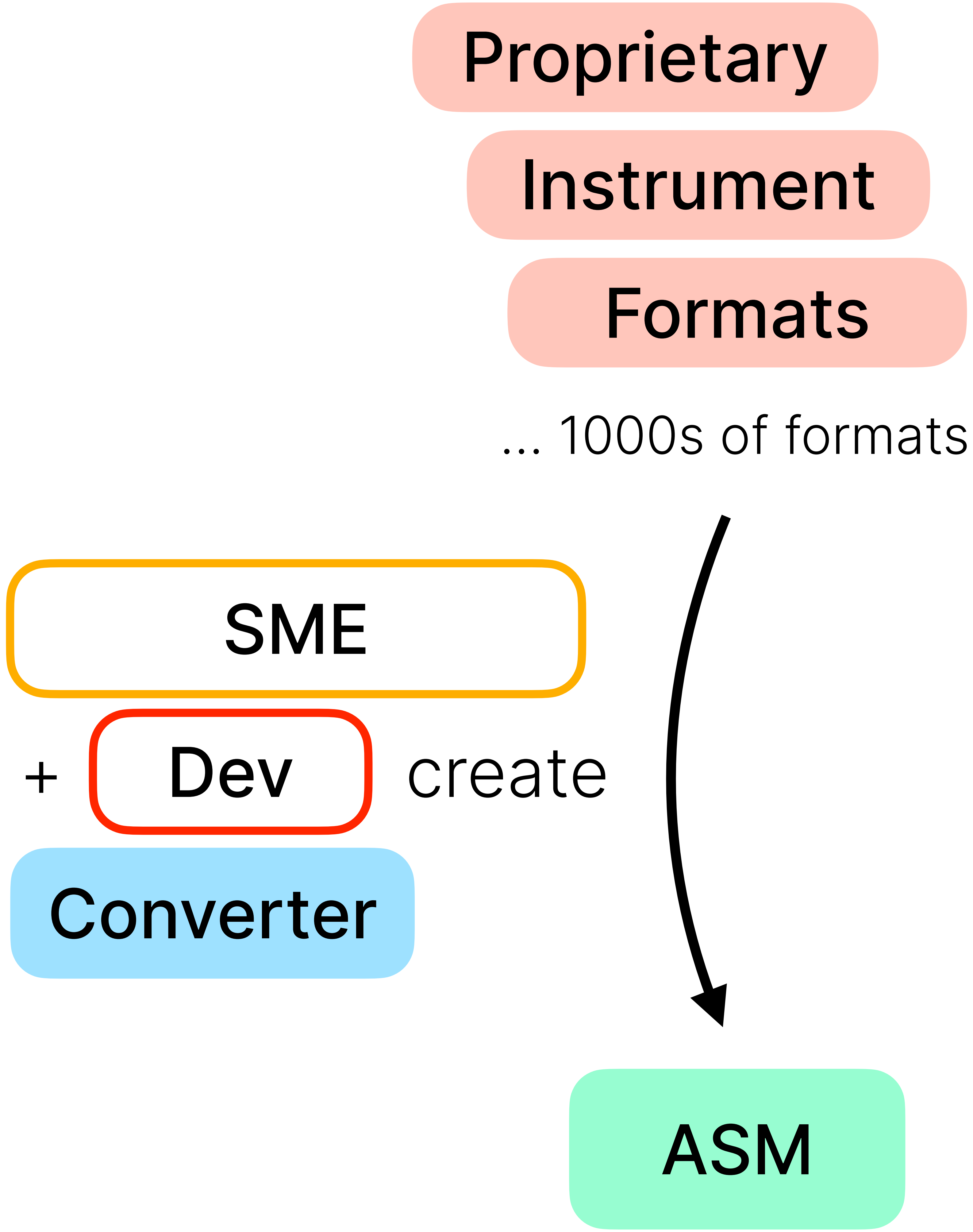
Dictionary

2-6 weeks
per language

**And repeat
100x!**

English

Status Quo



Proprietary

Instrument

Formats

... 1000s of formats

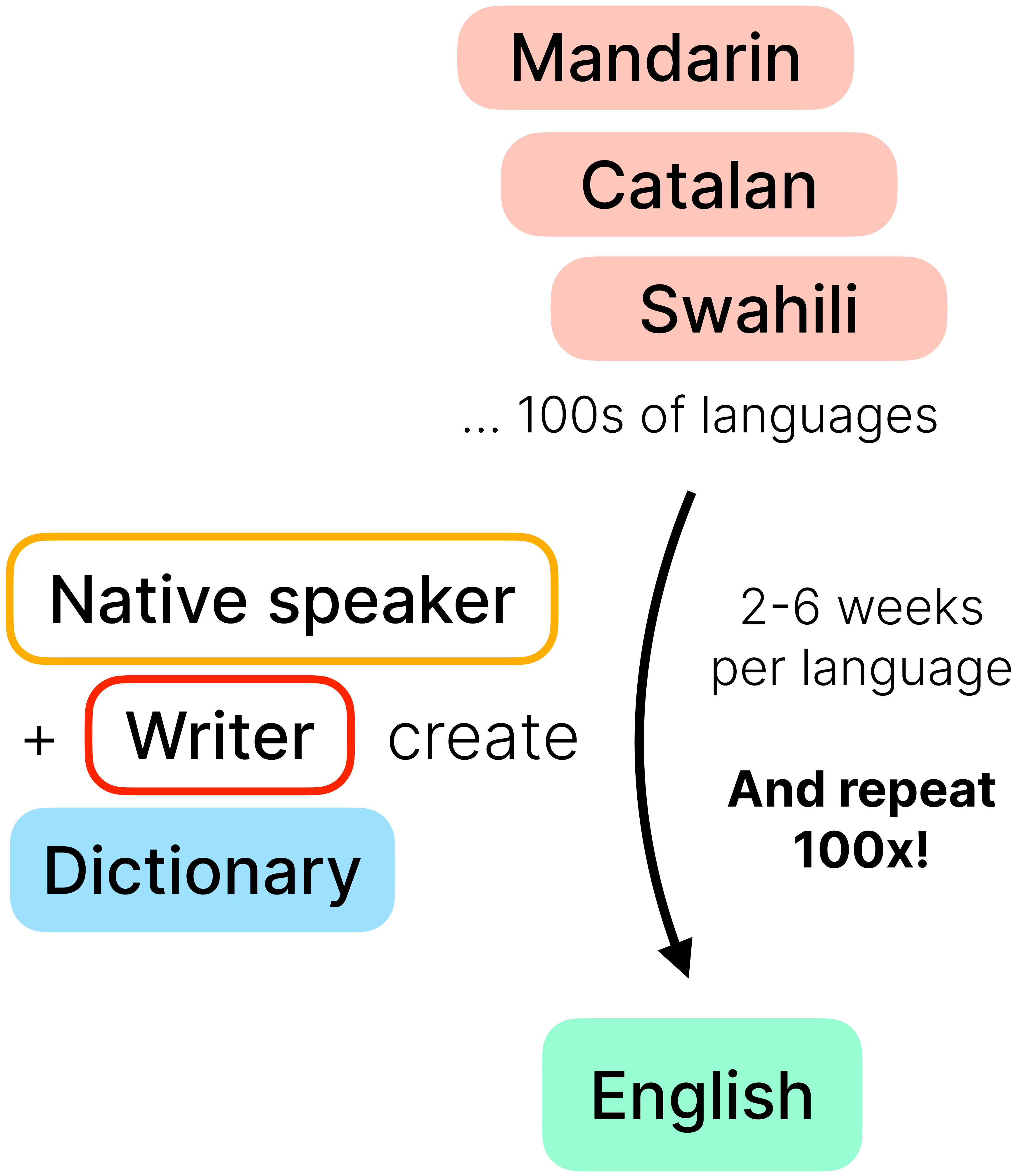
SME

+ Dev create

Converter

ASM

Analogy: Translation



Mandarin

Catalan

Swahili

... 100s of languages

Native speaker

+ Writer create

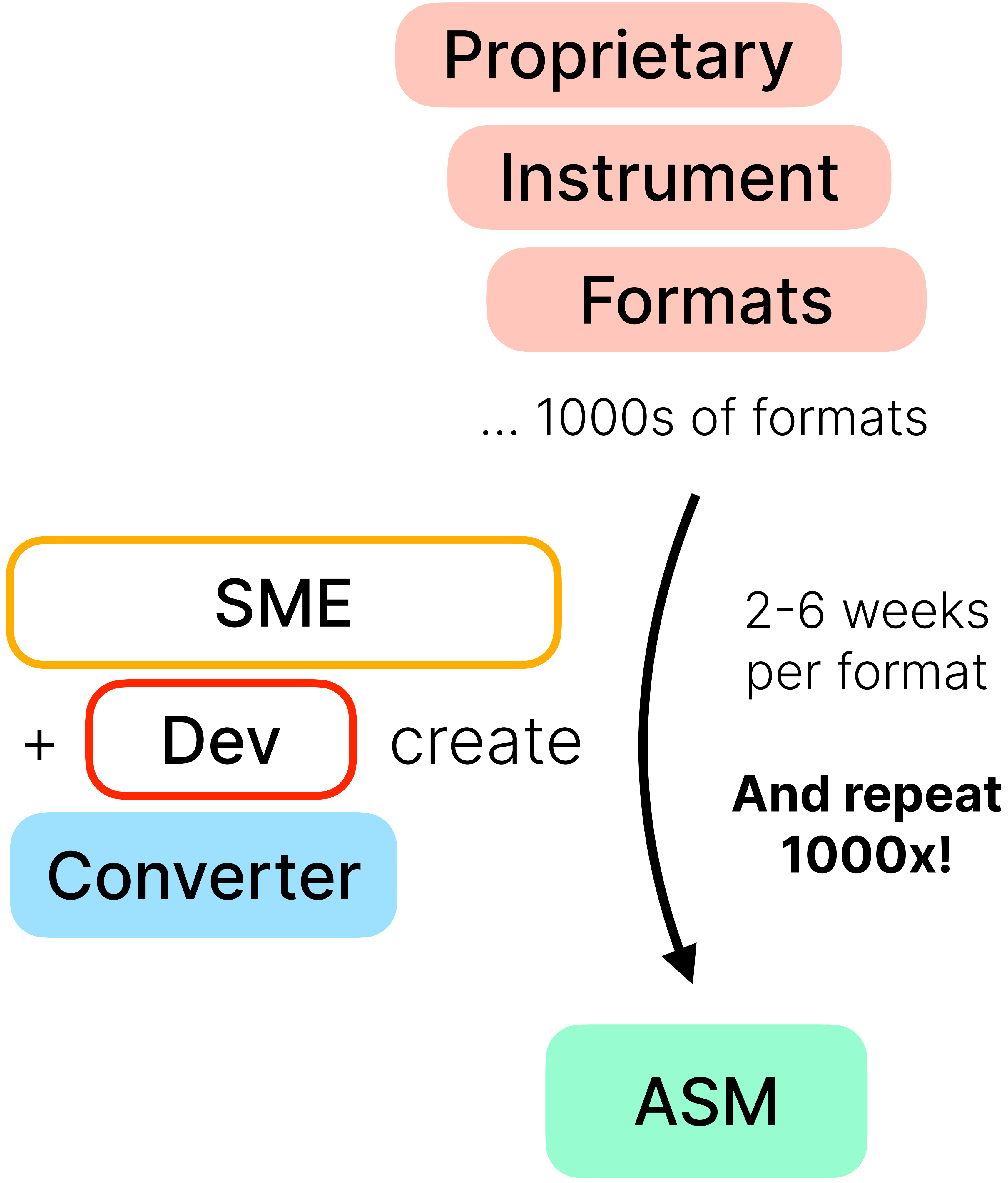
Dictionary

2-6 weeks
per language

**And repeat
100x!**

English

Status Quo



Proprietary

Instrument

Formats

... 1000s of formats

SME

+ Dev create

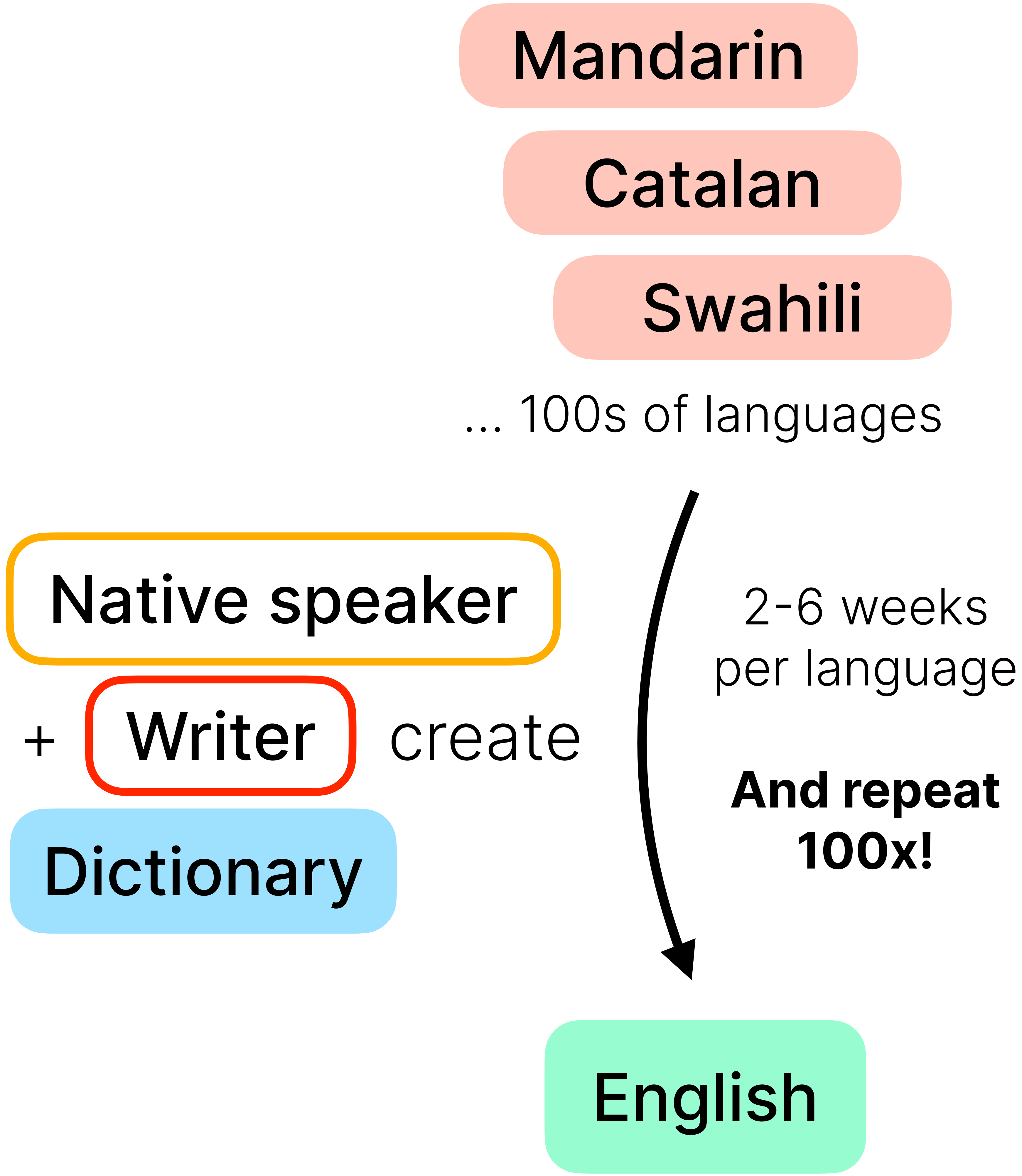
Converter

2-6 weeks
per format

**And repeat
1000x!**

ASM

Analogy: Translation



Mandarin

Catalan

Swahili

... 100s of languages

Native speaker

+ Writer create

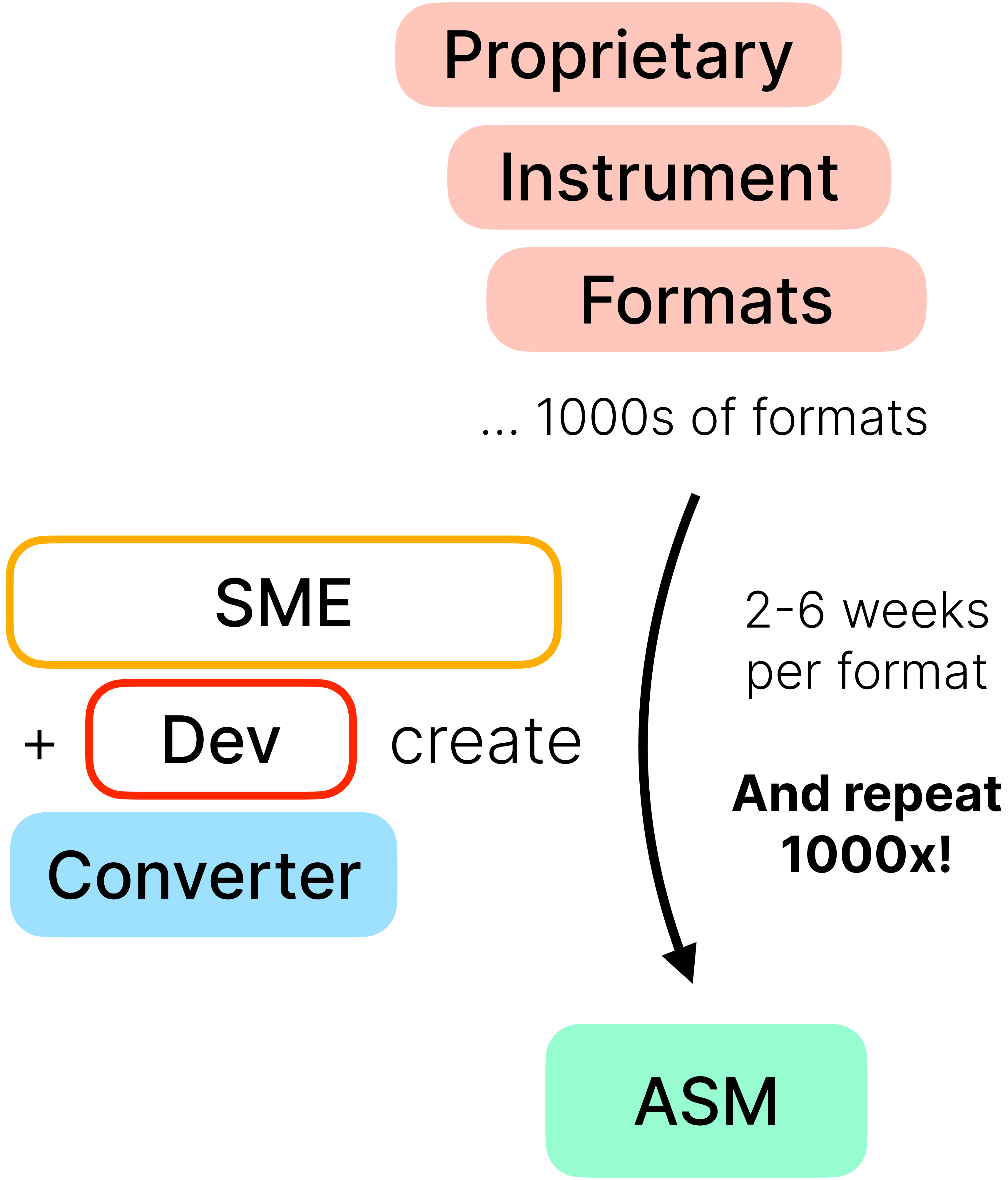
Dictionary

2-6 weeks
per language

**And repeat
100x!**

English

Status Quo



Proprietary

Instrument

Formats

... 1000s of formats

SME

+ Dev create

Converter

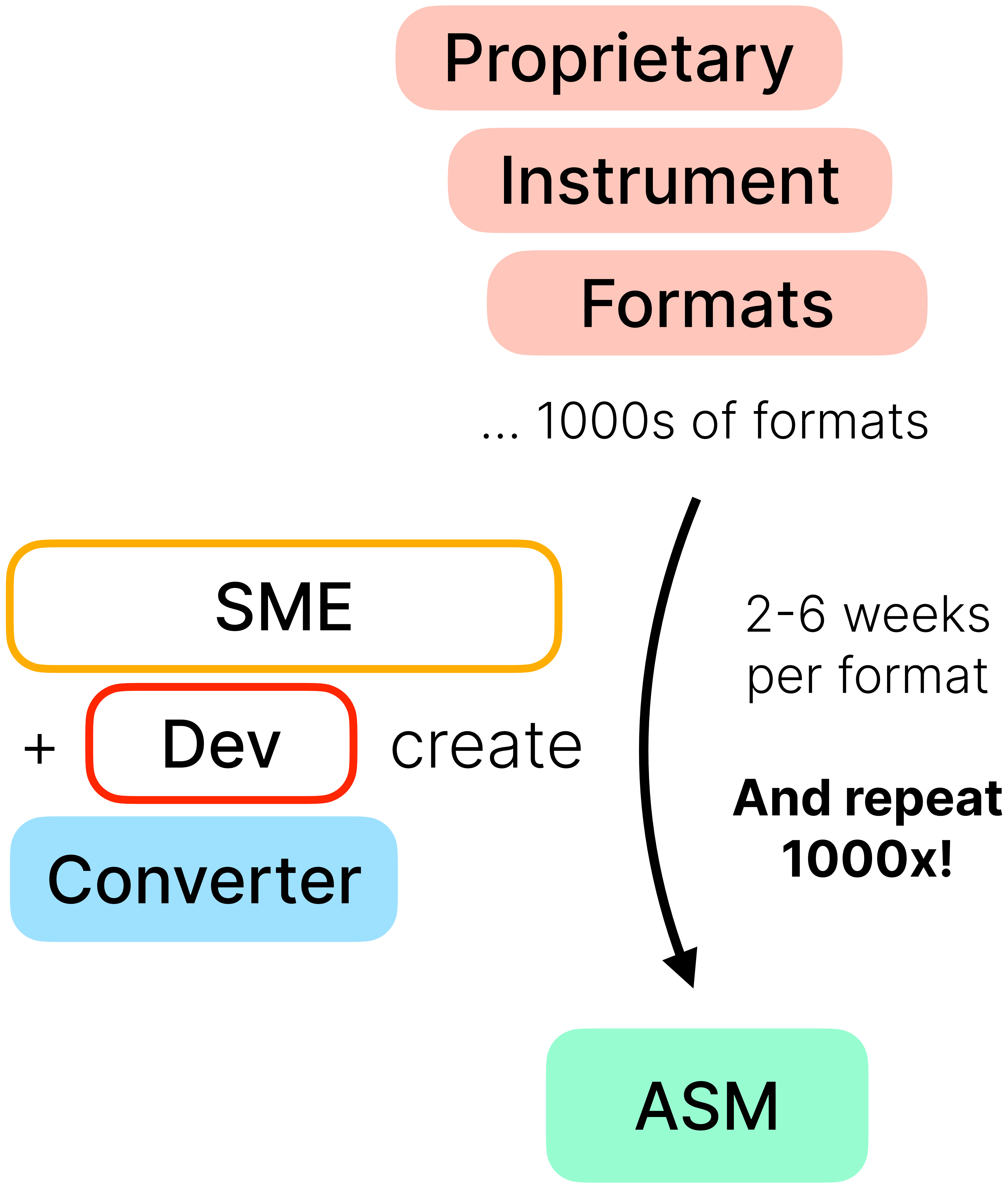
2-6 weeks
per format

**And repeat
1000x!**

ASM

Takes long time, resource-intensive

Status Quo



Proprietary

Instrument

Formats

... 1000s of formats

SME

+ Dev create

Converter

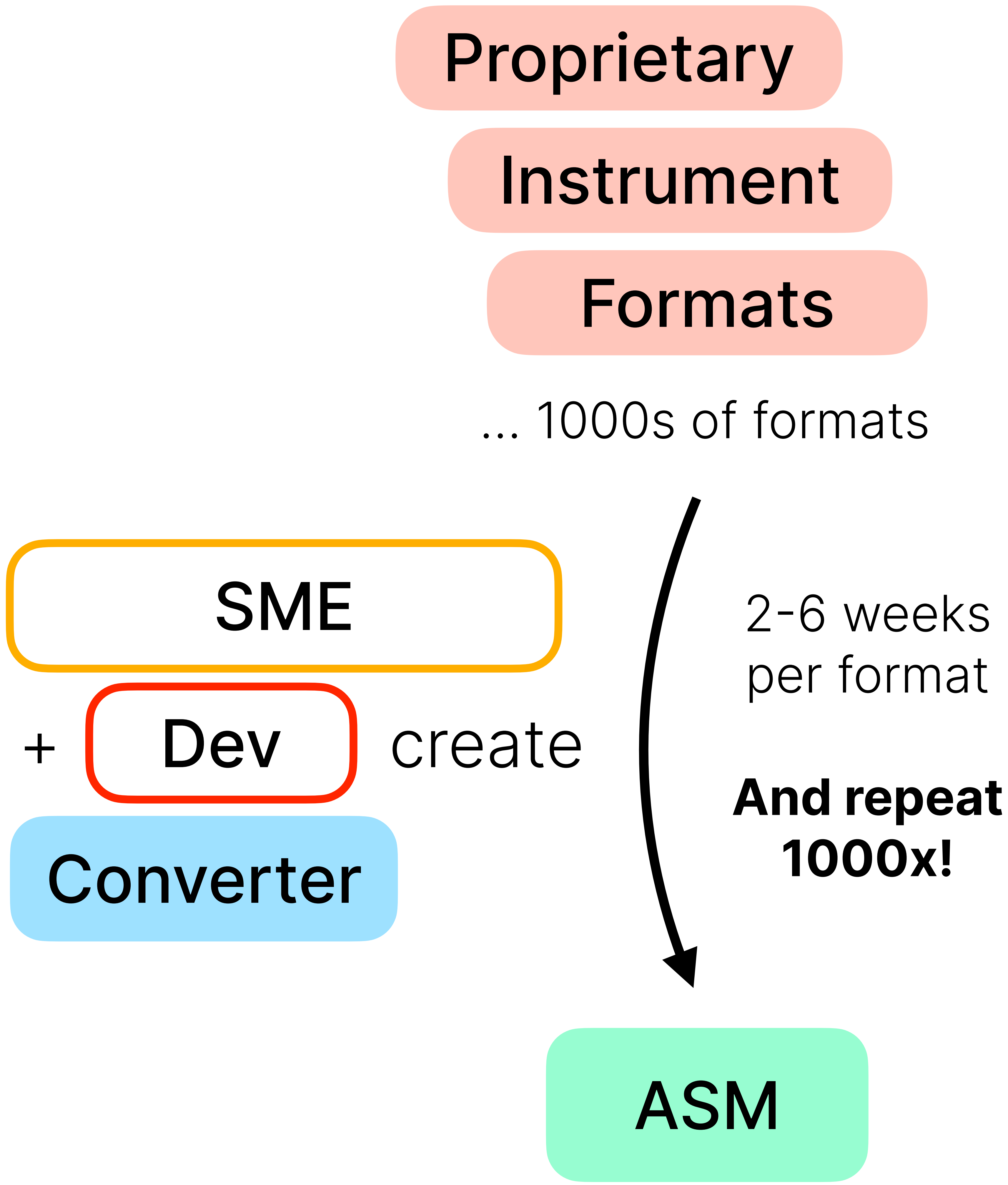
2-6 weeks
per format

**And repeat
1000x!**

ASM

Takes long time, resource-intensive

Status Quo



Proprietary

Instrument

Formats

... 1000s of formats

SME

+ Dev create

Converter

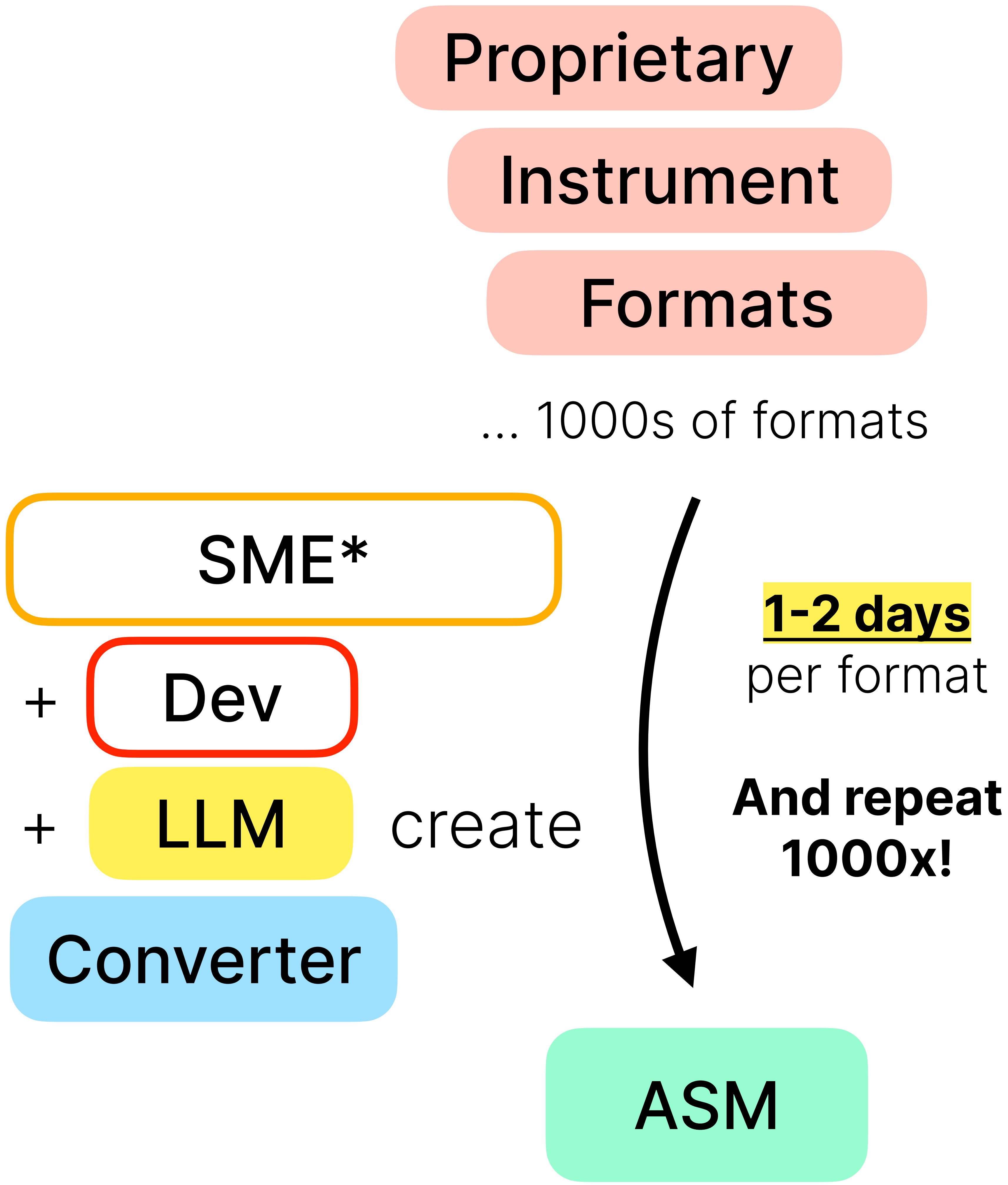
2-6 weeks
per format

**And repeat
1000x!**

ASM

Takes long time, resource-intensive

Our vision with LLMs



Proprietary

Instrument

Formats

... 1000s of formats

SME*

+ Dev

+ LLM create

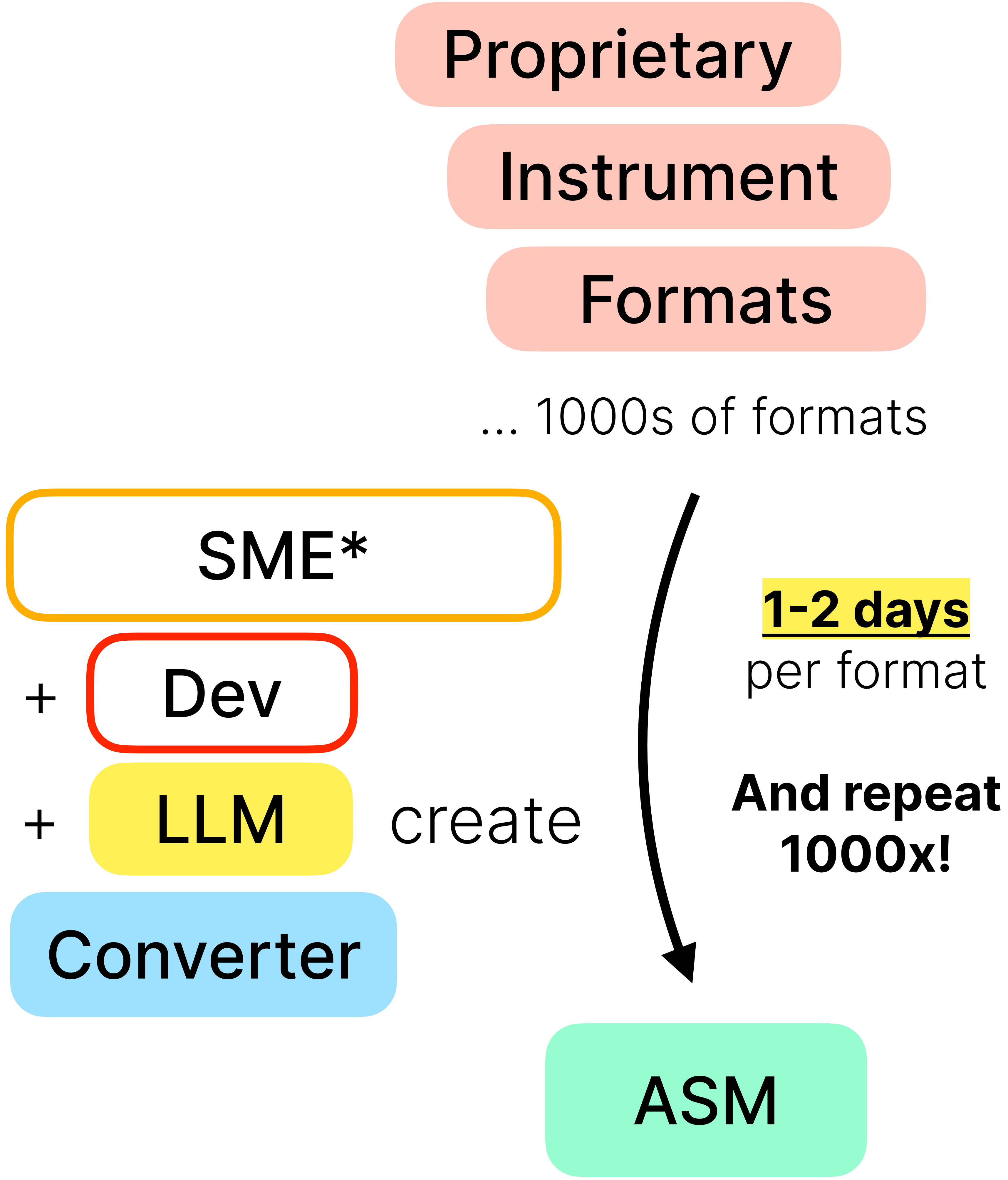
Converter

1-2 days
per format

**And repeat
1000x!**

ASM

Our vision with LLMs



Proprietary

Instrument

Formats

... 1000s of formats

SME*

+ Dev

+ LLM create

Converter

1-2 days

per format

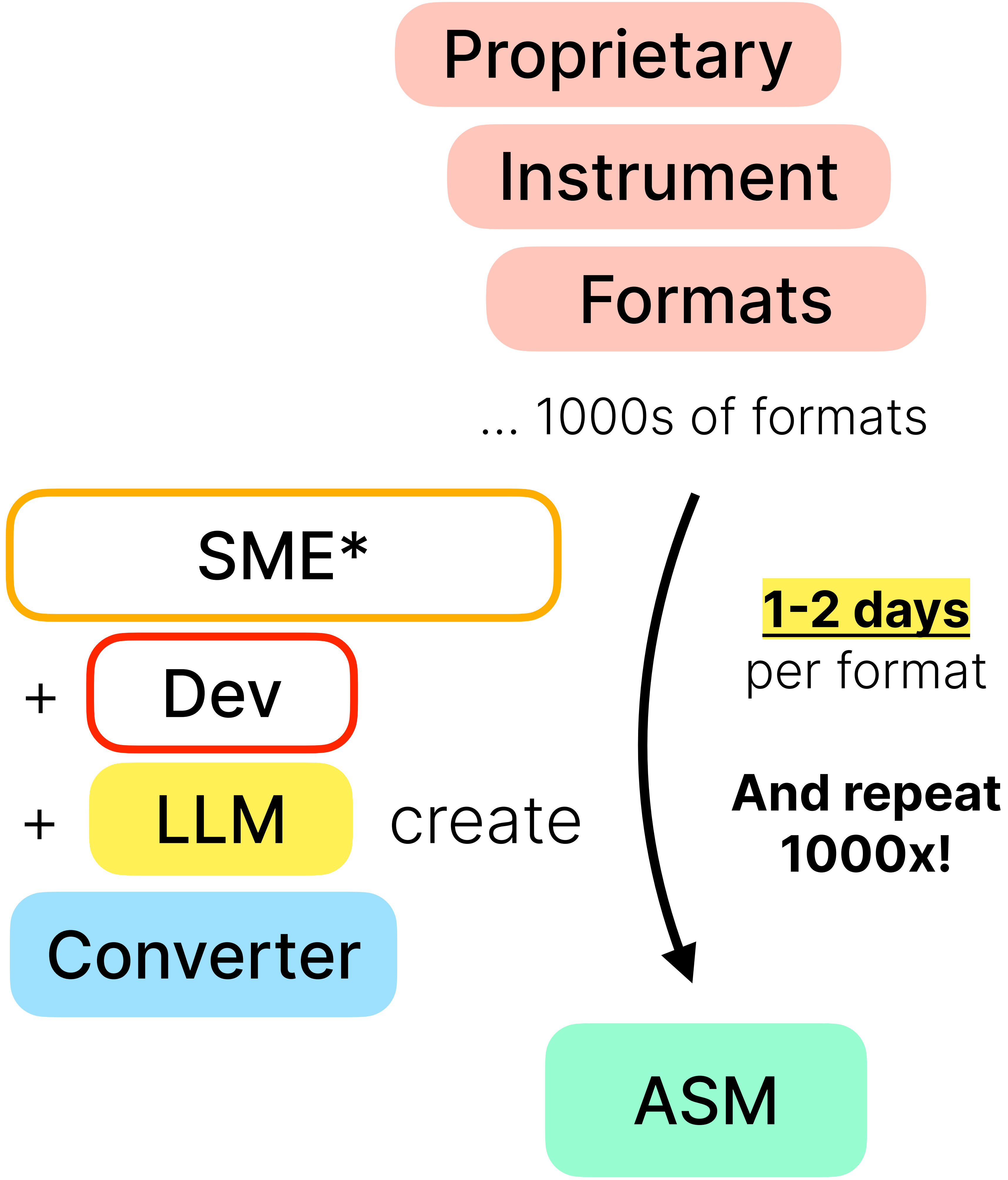
And repeat 1000x!

ASM

We used LLMs to build converters for 19 instruments

Open-source test set from Benching

Our vision with LLMs



Proprietary

Instrument

Formats

... 1000s of formats

SME*

+ Dev

+ LLM create

Converter

1-2 days per format

And repeat 1000x!

ASM

We used LLMs to build converters for 19 instruments

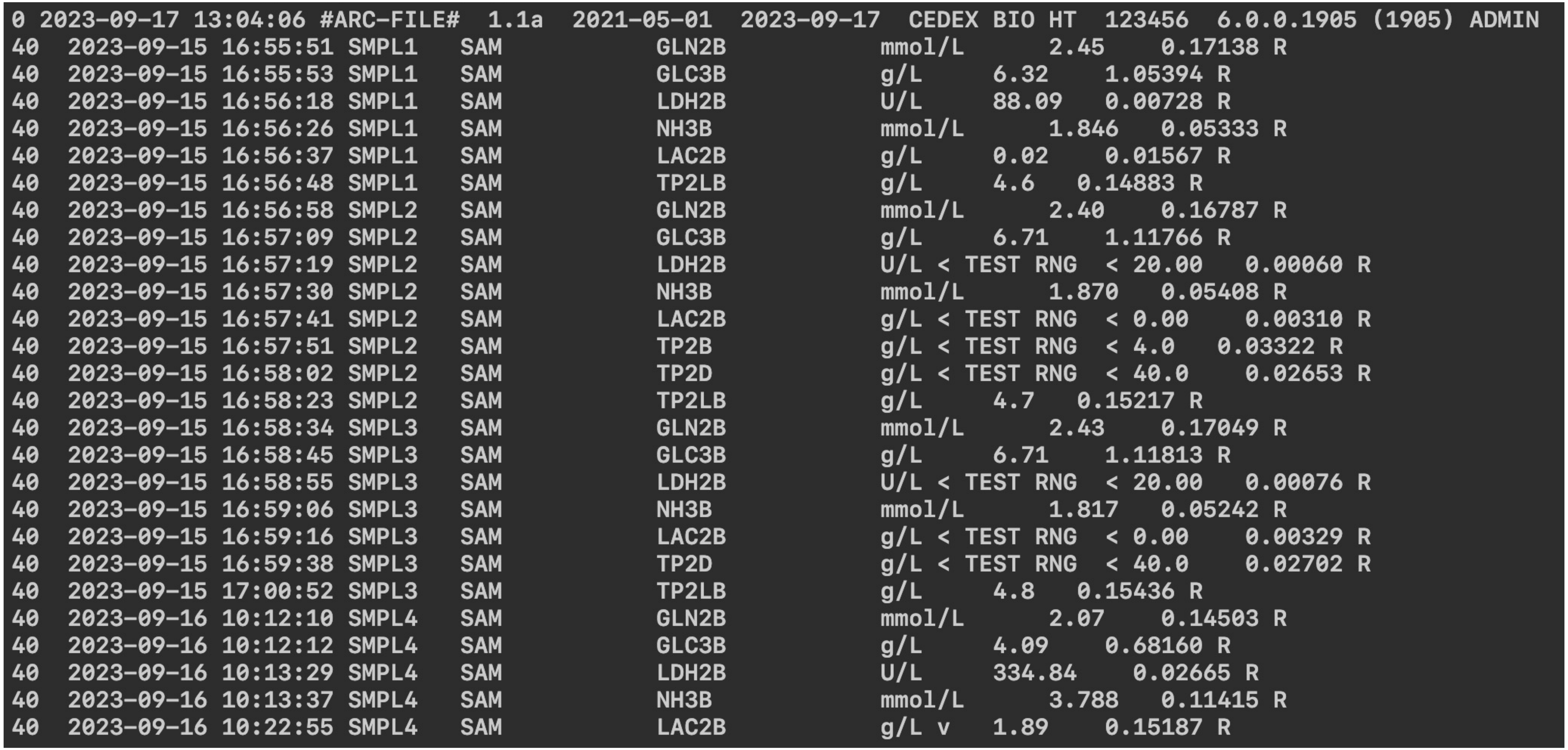
Open-source test set from Benching

Findings:

- 1-2 day turnaround end-to-end
- * SME involvement minimized

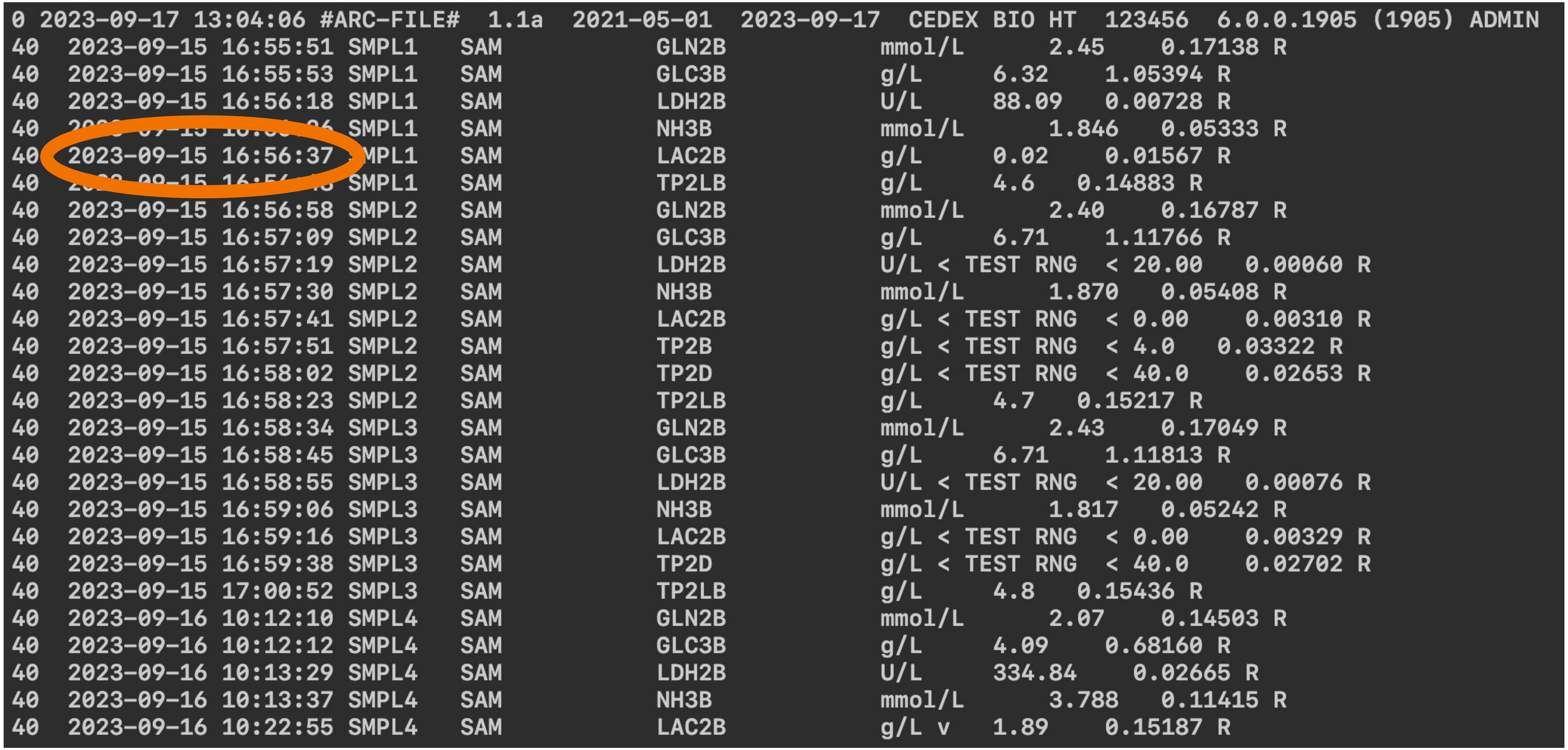
Only needed for validation, which is accelerated through LLM-specific tooling

Get familiar with the raw data from this cell culture analyzer...



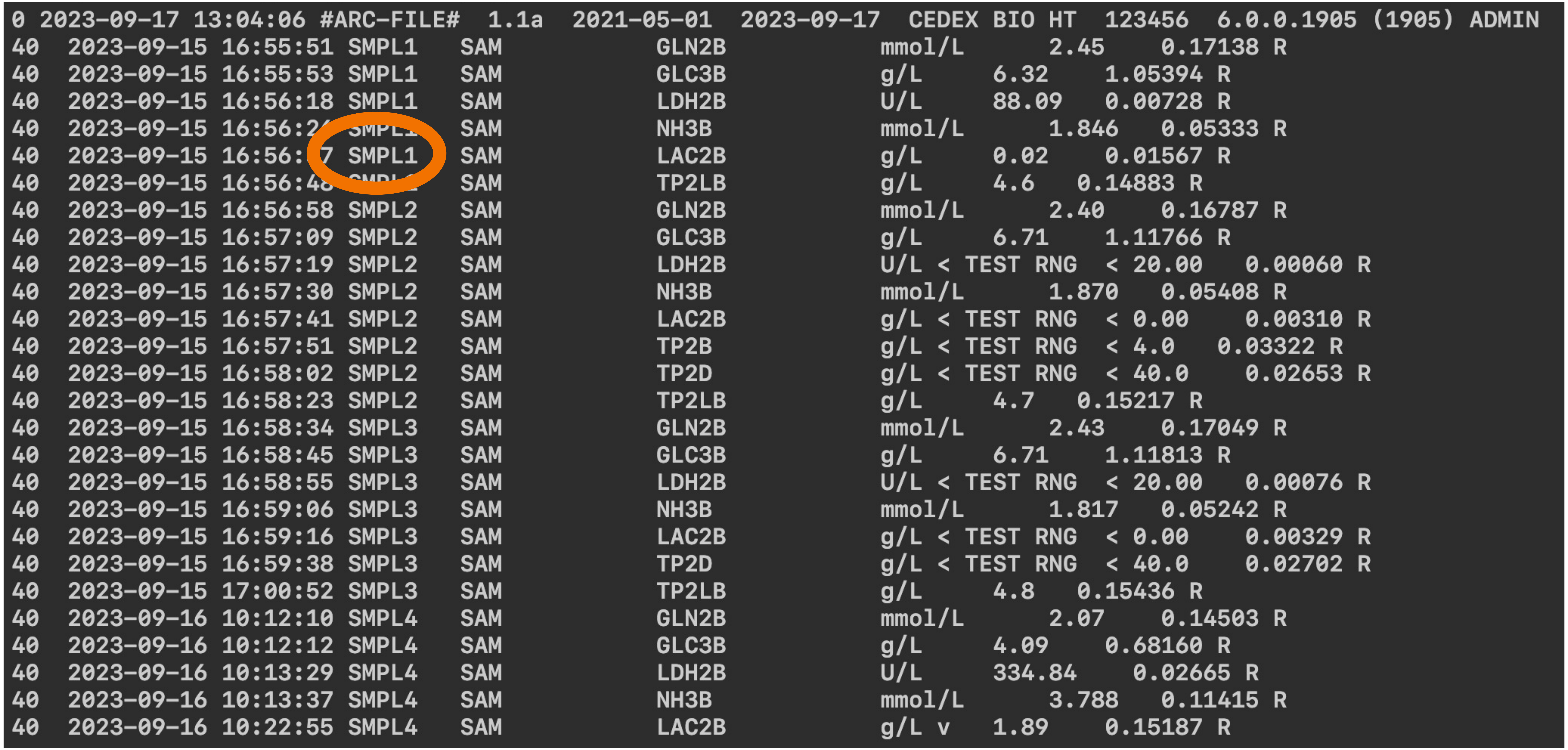
```
0 2023-09-17 13:04:06 #ARC-FILE#  1.1a  2021-05-01  2023-09-17  CEDEX BIO HT  123456  6.0.0.1905 (1905) ADMIN
40  2023-09-15 16:55:51 SMPL1   SAM           GLN2B           mmol/L      2.45    0.17138 R
40  2023-09-15 16:55:53 SMPL1   SAM           GLC3B           g/L     6.32    1.05394 R
40  2023-09-15 16:56:18 SMPL1   SAM           LDH2B           U/L     88.09   0.00728 R
40  2023-09-15 16:56:26 SMPL1   SAM           NH3B            mmol/L      1.846   0.05333 R
40  2023-09-15 16:56:37 SMPL1   SAM           LAC2B           g/L     0.02    0.01567 R
40  2023-09-15 16:56:48 SMPL1   SAM           TP2LB           g/L     4.6   0.14883 R
40  2023-09-15 16:56:58 SMPL2   SAM           GLN2B           mmol/L      2.40    0.16787 R
40  2023-09-15 16:57:09 SMPL2   SAM           GLC3B           g/L     6.71    1.11766 R
40  2023-09-15 16:57:19 SMPL2   SAM           LDH2B           U/L < TEST RNG  < 20.00   0.00060 R
40  2023-09-15 16:57:30 SMPL2   SAM           NH3B            mmol/L      1.870   0.05408 R
40  2023-09-15 16:57:41 SMPL2   SAM           LAC2B           g/L < TEST RNG  < 0.00    0.00310 R
40  2023-09-15 16:57:51 SMPL2   SAM           TP2B            g/L < TEST RNG  < 4.0   0.03322 R
40  2023-09-15 16:58:02 SMPL2   SAM           TP2D            g/L < TEST RNG  < 40.0    0.02653 R
40  2023-09-15 16:58:23 SMPL2   SAM           TP2LB           g/L     4.7   0.15217 R
40  2023-09-15 16:58:34 SMPL3   SAM           GLN2B           mmol/L      2.43    0.17049 R
40  2023-09-15 16:58:45 SMPL3   SAM           GLC3B           g/L     6.71    1.11813 R
40  2023-09-15 16:58:55 SMPL3   SAM           LDH2B           U/L < TEST RNG  < 20.00   0.00076 R
40  2023-09-15 16:59:06 SMPL3   SAM           NH3B            mmol/L      1.817   0.05242 R
40  2023-09-15 16:59:16 SMPL3   SAM           LAC2B           g/L < TEST RNG  < 0.00    0.00329 R
40  2023-09-15 16:59:38 SMPL3   SAM           TP2D            g/L < TEST RNG  < 40.0    0.02702 R
40  2023-09-15 17:00:52 SMPL3   SAM           TP2LB           g/L     4.8   0.15436 R
40  2023-09-16 10:12:10 SMPL4   SAM           GLN2B           mmol/L      2.07    0.14503 R
40  2023-09-16 10:12:12 SMPL4   SAM           GLC3B           g/L     4.09    0.68160 R
40  2023-09-16 10:13:29 SMPL4   SAM           LDH2B           U/L     334.84    0.02665 R
40  2023-09-16 10:13:37 SMPL4   SAM           NH3B            mmol/L      3.788   0.11415 R
40  2023-09-16 10:22:55 SMPL4   SAM           LAC2B           g/L v   1.89    0.15187 R
```

Get familiar with the raw data from this cell culture analyzer...



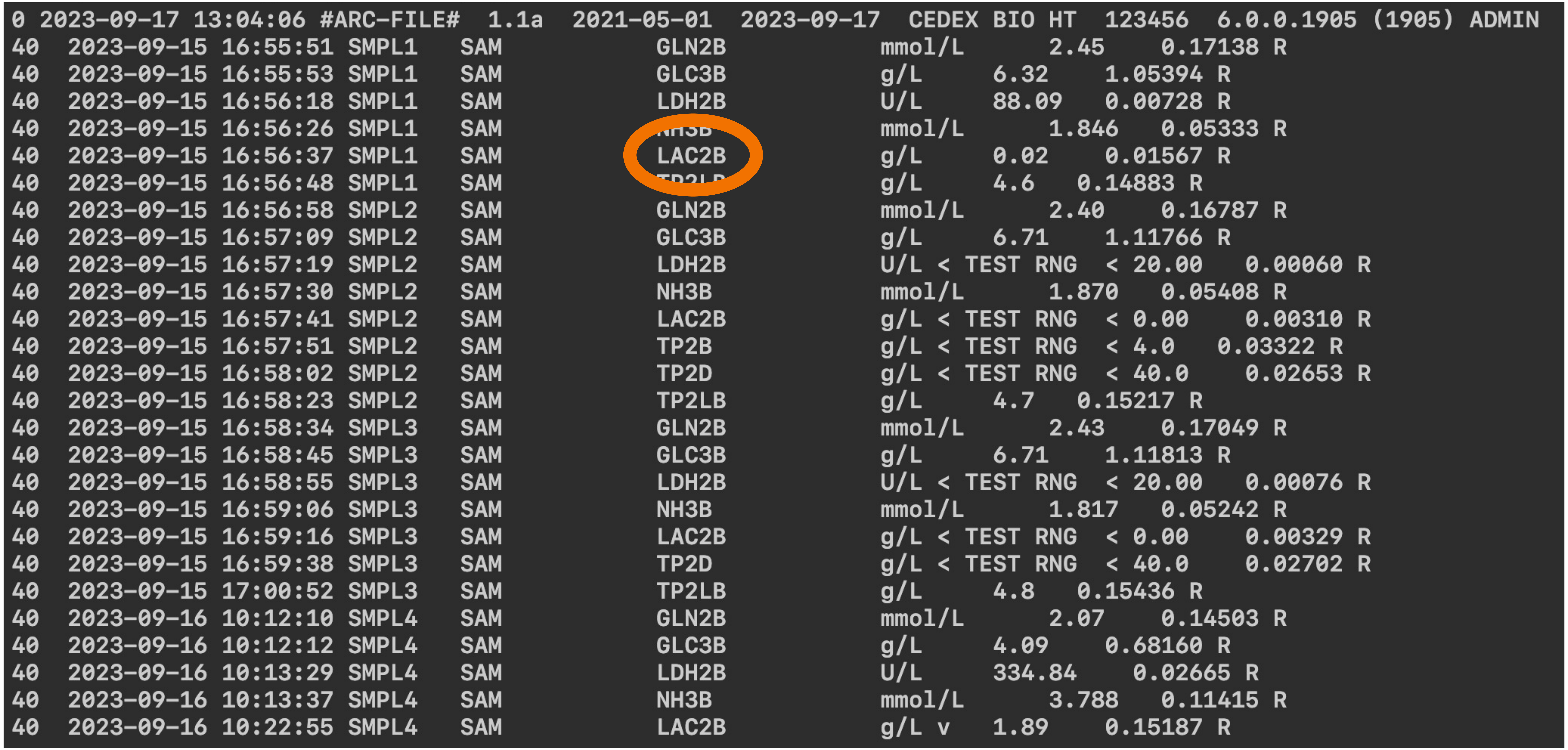
```
0 2023-09-17 13:04:06 #ARC-FILE#  1.1a  2021-05-01  2023-09-17  CEDEX BIO HT  123456  6.0.0.1905 (1905) ADMIN
40  2023-09-15 16:55:51 SMPL1   SAM           GLN2B           mmol/L      2.45    0.17138 R
40  2023-09-15 16:55:53 SMPL1   SAM           GLC3B           g/L     6.32    1.05394 R
40  2023-09-15 16:56:18 SMPL1   SAM           LDH2B           U/L     88.09   0.00728 R
40  2023-09-15 16:56:26 SMPL1   SAM           NH3B            mmol/L      1.846   0.05333 R
40  2023-09-15 16:56:37 SMPL1   SAM           LAC2B           g/L     0.02    0.01567 R
40  2023-09-15 16:56:48 SMPL1   SAM           TP2LB           g/L     4.6   0.14883 R
40  2023-09-15 16:56:58 SMPL2   SAM           GLN2B           mmol/L      2.40    0.16787 R
40  2023-09-15 16:57:09 SMPL2   SAM           GLC3B           g/L     6.71    1.11766 R
40  2023-09-15 16:57:19 SMPL2   SAM           LDH2B           U/L < TEST RNG  < 20.00   0.00060 R
40  2023-09-15 16:57:30 SMPL2   SAM           NH3B            mmol/L      1.870   0.05408 R
40  2023-09-15 16:57:41 SMPL2   SAM           LAC2B           g/L < TEST RNG  < 0.00    0.00310 R
40  2023-09-15 16:57:51 SMPL2   SAM           TP2B            g/L < TEST RNG  < 4.0   0.03322 R
40  2023-09-15 16:58:02 SMPL2   SAM           TP2D            g/L < TEST RNG  < 40.0    0.02653 R
40  2023-09-15 16:58:23 SMPL2   SAM           TP2LB           g/L     4.7   0.15217 R
40  2023-09-15 16:58:34 SMPL3   SAM           GLN2B           mmol/L      2.43    0.17049 R
40  2023-09-15 16:58:45 SMPL3   SAM           GLC3B           g/L     6.71    1.11813 R
40  2023-09-15 16:58:55 SMPL3   SAM           LDH2B           U/L < TEST RNG  < 20.00   0.00076 R
40  2023-09-15 16:59:06 SMPL3   SAM           NH3B            mmol/L      1.817   0.05242 R
40  2023-09-15 16:59:16 SMPL3   SAM           LAC2B           g/L < TEST RNG  < 0.00    0.00329 R
40  2023-09-15 16:59:38 SMPL3   SAM           TP2D            g/L < TEST RNG  < 40.0    0.02702 R
40  2023-09-15 17:00:52 SMPL3   SAM           TP2LB           g/L     4.8   0.15436 R
40  2023-09-16 10:12:10 SMPL4   SAM           GLN2B           mmol/L      2.07    0.14503 R
40  2023-09-16 10:12:12 SMPL4   SAM           GLC3B           g/L     4.09    0.68160 R
40  2023-09-16 10:13:29 SMPL4   SAM           LDH2B           U/L     334.84    0.02665 R
40  2023-09-16 10:13:37 SMPL4   SAM           NH3B            mmol/L      3.788   0.11415 R
40  2023-09-16 10:22:55 SMPL4   SAM           LAC2B           g/L v   1.89    0.15187 R
```

Get familiar with the raw data from this cell culture analyzer...



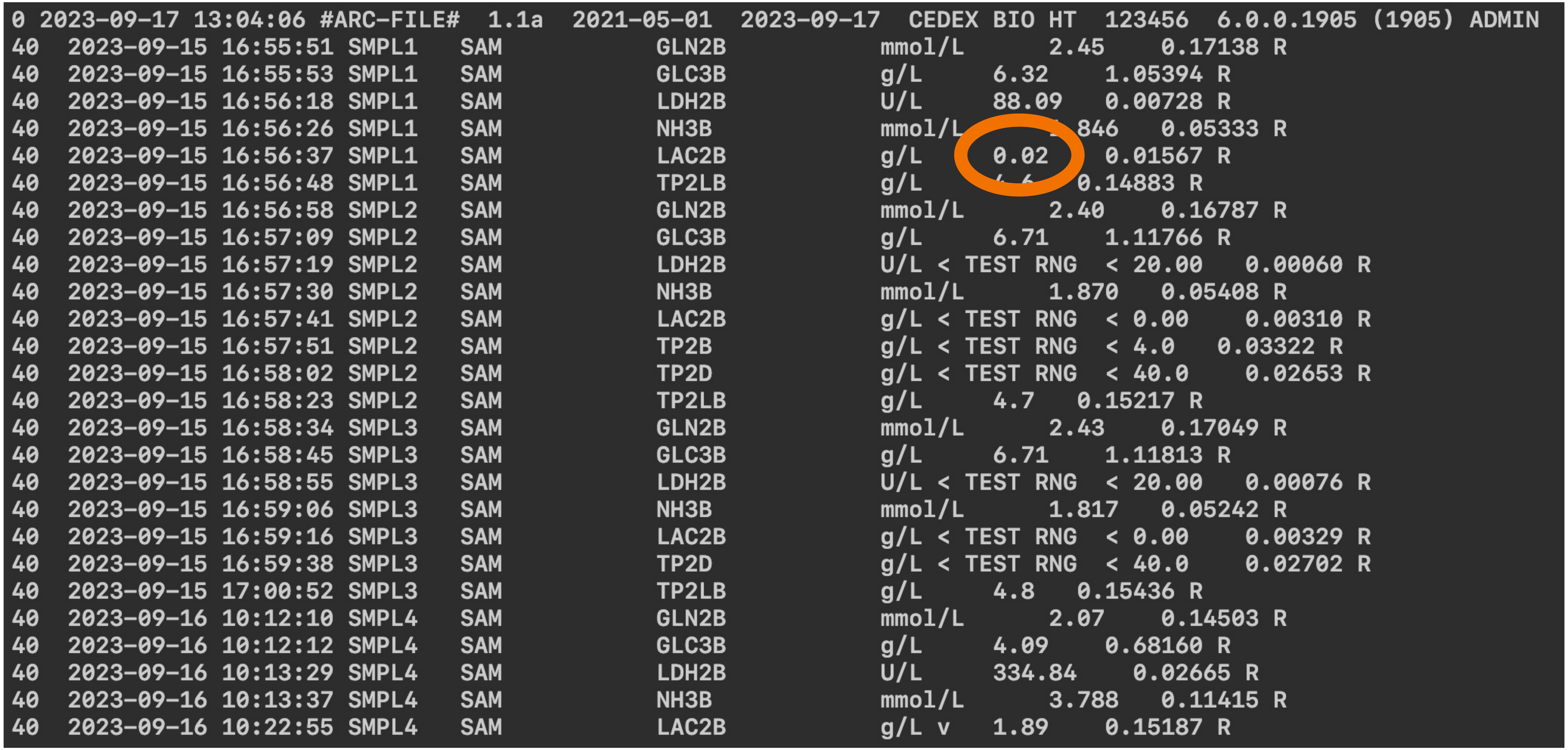
```
0 2023-09-17 13:04:06 #ARC-FILE#  1.1a  2021-05-01  2023-09-17  CEDEX BIO HT  123456  6.0.0.1905 (1905) ADMIN
40  2023-09-15 16:55:51 SMPL1   SAM           GLN2B          mmol/L      2.45    0.17138 R
40  2023-09-15 16:55:53 SMPL1   SAM           GLC3B          g/L     6.32    1.05394 R
40  2023-09-15 16:56:18 SMPL1   SAM           LDH2B          U/L     88.09   0.00728 R
40  2023-09-15 16:56:24 SMPL1   SAM           NH3B           mmol/L      1.846   0.05333 R
40  2023-09-15 16:56:37 SMPL1   SAM           LAC2B          g/L     0.02    0.01567 R
40  2023-09-15 16:56:48 SMPL1   SAM           TP2LB          g/L     4.6    0.14883 R
40  2023-09-15 16:56:58 SMPL2   SAM           GLN2B          mmol/L      2.40    0.16787 R
40  2023-09-15 16:57:09 SMPL2   SAM           GLC3B          g/L     6.71    1.11766 R
40  2023-09-15 16:57:19 SMPL2   SAM           LDH2B          U/L < TEST RNG  < 20.00   0.00060 R
40  2023-09-15 16:57:30 SMPL2   SAM           NH3B           mmol/L      1.870   0.05408 R
40  2023-09-15 16:57:41 SMPL2   SAM           LAC2B          g/L < TEST RNG  < 0.00    0.00310 R
40  2023-09-15 16:57:51 SMPL2   SAM           TP2B           g/L < TEST RNG  < 4.0   0.03322 R
40  2023-09-15 16:58:02 SMPL2   SAM           TP2D           g/L < TEST RNG  < 40.0    0.02653 R
40  2023-09-15 16:58:23 SMPL2   SAM           TP2LB          g/L     4.7    0.15217 R
40  2023-09-15 16:58:34 SMPL3   SAM           GLN2B          mmol/L      2.43    0.17049 R
40  2023-09-15 16:58:45 SMPL3   SAM           GLC3B          g/L     6.71    1.11813 R
40  2023-09-15 16:58:55 SMPL3   SAM           LDH2B          U/L < TEST RNG  < 20.00   0.00076 R
40  2023-09-15 16:59:06 SMPL3   SAM           NH3B           mmol/L      1.817   0.05242 R
40  2023-09-15 16:59:16 SMPL3   SAM           LAC2B          g/L < TEST RNG  < 0.00    0.00329 R
40  2023-09-15 16:59:38 SMPL3   SAM           TP2D           g/L < TEST RNG  < 40.0    0.02702 R
40  2023-09-15 17:00:52 SMPL3   SAM           TP2LB          g/L     4.8    0.15436 R
40  2023-09-16 10:12:10 SMPL4   SAM           GLN2B          mmol/L      2.07    0.14503 R
40  2023-09-16 10:12:12 SMPL4   SAM           GLC3B          g/L     4.09    0.68160 R
40  2023-09-16 10:13:29 SMPL4   SAM           LDH2B          U/L     334.84    0.02665 R
40  2023-09-16 10:13:37 SMPL4   SAM           NH3B           mmol/L      3.788   0.11415 R
40  2023-09-16 10:22:55 SMPL4   SAM           LAC2B          g/L v   1.89    0.15187 R
```

Get familiar with the raw data
from this cell culture analyzer...



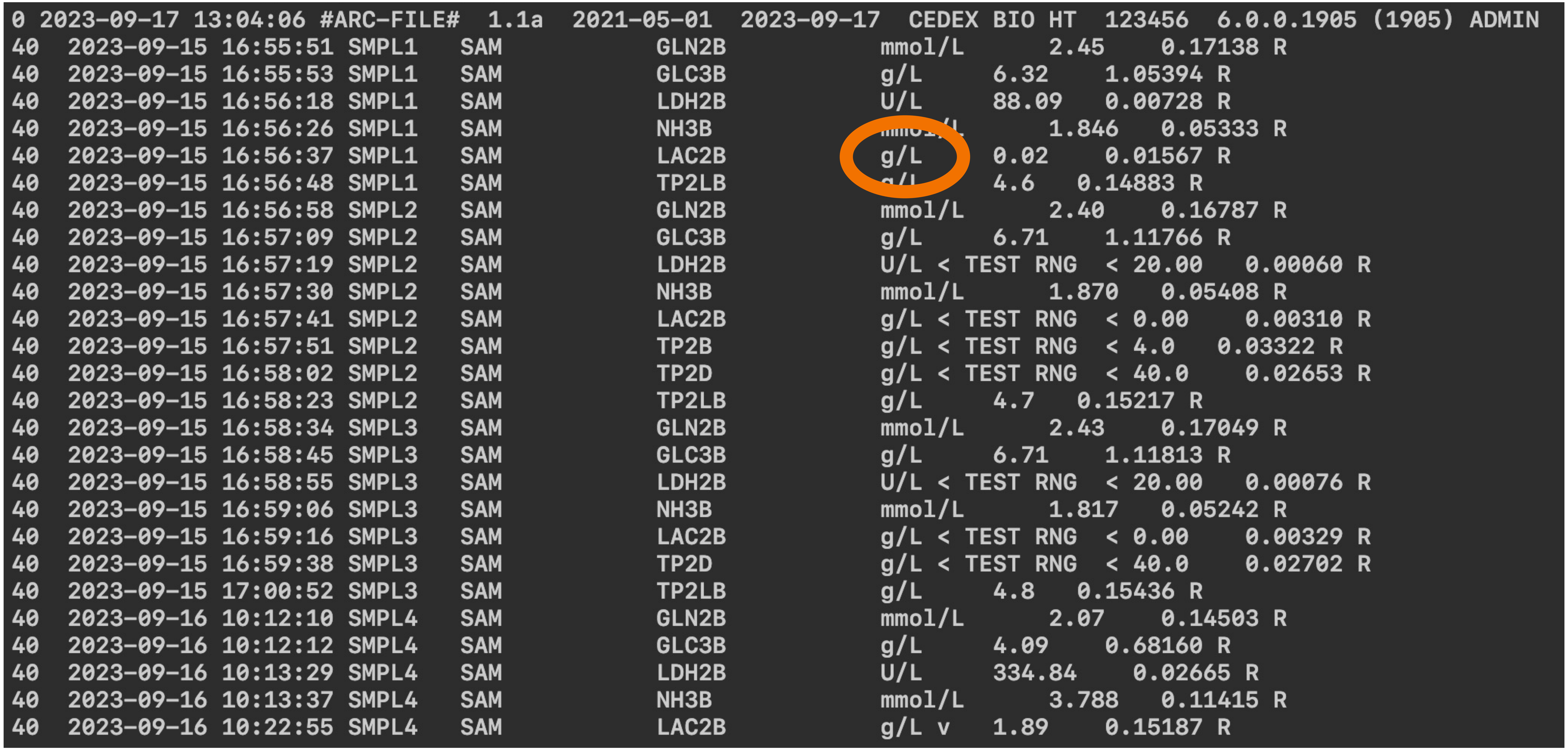
```
0 2023-09-17 13:04:06 #ARC-FILE#  1.1a  2021-05-01  2023-09-17  CEDEX BIO HT  123456  6.0.0.1905 (1905) ADMIN
40  2023-09-15 16:55:51 SMPL1   SAM           GLN2B           mmol/L      2.45    0.17138 R
40  2023-09-15 16:55:53 SMPL1   SAM           GLC3B           g/L     6.32    1.05394 R
40  2023-09-15 16:56:18 SMPL1   SAM           LDH2B           U/L     88.09   0.00728 R
40  2023-09-15 16:56:26 SMPL1   SAM           NH3B            mmol/L      1.846   0.05333 R
40  2023-09-15 16:56:37 SMPL1   SAM           LAC2B           g/L     0.02    0.01567 R
40  2023-09-15 16:56:48 SMPL1   SAM           TP2LB           g/L     4.6   0.14883 R
40  2023-09-15 16:56:58 SMPL2   SAM           GLN2B           mmol/L      2.40    0.16787 R
40  2023-09-15 16:57:09 SMPL2   SAM           GLC3B           g/L     6.71    1.11766 R
40  2023-09-15 16:57:19 SMPL2   SAM           LDH2B           U/L < TEST RNG  < 20.00   0.00060 R
40  2023-09-15 16:57:30 SMPL2   SAM           NH3B            mmol/L      1.870   0.05408 R
40  2023-09-15 16:57:41 SMPL2   SAM           LAC2B           g/L < TEST RNG  < 0.00    0.00310 R
40  2023-09-15 16:57:51 SMPL2   SAM           TP2B            g/L < TEST RNG  < 4.0   0.03322 R
40  2023-09-15 16:58:02 SMPL2   SAM           TP2D            g/L < TEST RNG  < 40.0    0.02653 R
40  2023-09-15 16:58:23 SMPL2   SAM           TP2LB           g/L     4.7   0.15217 R
40  2023-09-15 16:58:34 SMPL3   SAM           GLN2B           mmol/L      2.43    0.17049 R
40  2023-09-15 16:58:45 SMPL3   SAM           GLC3B           g/L     6.71    1.11813 R
40  2023-09-15 16:58:55 SMPL3   SAM           LDH2B           U/L < TEST RNG  < 20.00   0.00076 R
40  2023-09-15 16:59:06 SMPL3   SAM           NH3B            mmol/L      1.817   0.05242 R
40  2023-09-15 16:59:16 SMPL3   SAM           LAC2B           g/L < TEST RNG  < 0.00    0.00329 R
40  2023-09-15 16:59:38 SMPL3   SAM           TP2D            g/L < TEST RNG  < 40.0    0.02702 R
40  2023-09-15 17:00:52 SMPL3   SAM           TP2LB           g/L     4.8   0.15436 R
40  2023-09-16 10:12:10 SMPL4   SAM           GLN2B           mmol/L      2.07    0.14503 R
40  2023-09-16 10:12:12 SMPL4   SAM           GLC3B           g/L     4.09    0.68160 R
40  2023-09-16 10:13:29 SMPL4   SAM           LDH2B           U/L     334.84    0.02665 R
40  2023-09-16 10:13:37 SMPL4   SAM           NH3B            mmol/L      3.788   0.11415 R
40  2023-09-16 10:22:55 SMPL4   SAM           LAC2B           g/L v   1.89    0.15187 R
```

Get familiar with the raw data from this cell culture analyzer...



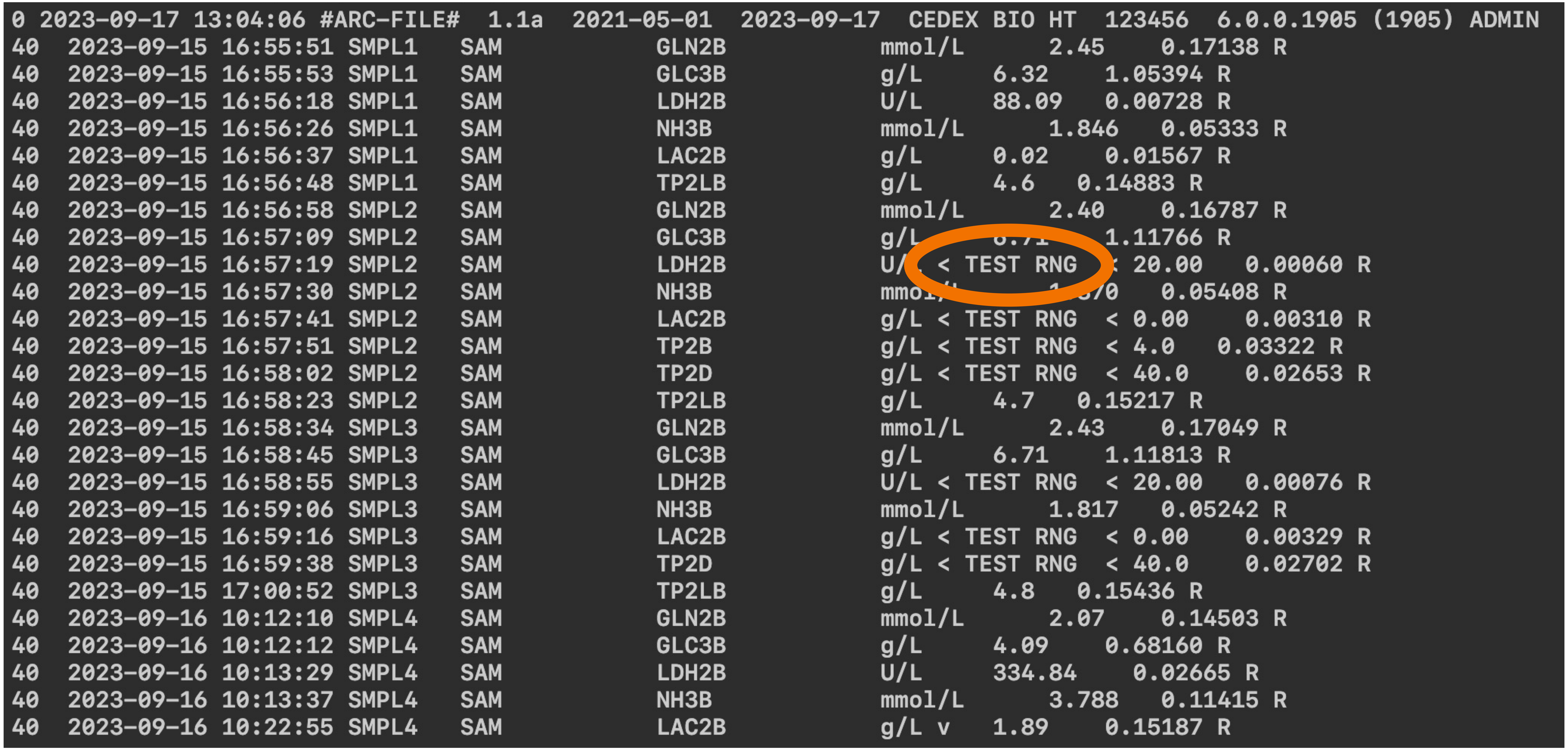
```
0 2023-09-17 13:04:06 #ARC-FILE#  1.1a  2021-05-01  2023-09-17  CEDEX BIO HT  123456  6.0.0.1905 (1905)  ADMIN
40  2023-09-15 16:55:51 SMPL1   SAM           GLN2B           mmol/L      2.45    0.17138 R
40  2023-09-15 16:55:53 SMPL1   SAM           GLC3B           g/L     6.32    1.05394 R
40  2023-09-15 16:56:18 SMPL1   SAM           LDH2B           U/L     88.09   0.00728 R
40  2023-09-15 16:56:26 SMPL1   SAM           NH3B            mmol/L      1.846   0.05333 R
40  2023-09-15 16:56:37 SMPL1   SAM           LAC2B           g/L     0.02    0.01567 R
40  2023-09-15 16:56:48 SMPL1   SAM           TP2LB           g/L     4.6   0.14883 R
40  2023-09-15 16:56:58 SMPL2   SAM           GLN2B           mmol/L      2.40    0.16787 R
40  2023-09-15 16:57:09 SMPL2   SAM           GLC3B           g/L     6.71    1.11766 R
40  2023-09-15 16:57:19 SMPL2   SAM           LDH2B           U/L < TEST RNG  < 20.00   0.00060 R
40  2023-09-15 16:57:30 SMPL2   SAM           NH3B            mmol/L      1.870   0.05408 R
40  2023-09-15 16:57:41 SMPL2   SAM           LAC2B           g/L < TEST RNG  < 0.00    0.00310 R
40  2023-09-15 16:57:51 SMPL2   SAM           TP2B            g/L < TEST RNG  < 4.0   0.03322 R
40  2023-09-15 16:58:02 SMPL2   SAM           TP2D            g/L < TEST RNG  < 40.0    0.02653 R
40  2023-09-15 16:58:23 SMPL2   SAM           TP2LB           g/L     4.7   0.15217 R
40  2023-09-15 16:58:34 SMPL3   SAM           GLN2B           mmol/L      2.43    0.17049 R
40  2023-09-15 16:58:45 SMPL3   SAM           GLC3B           g/L     6.71    1.11813 R
40  2023-09-15 16:58:55 SMPL3   SAM           LDH2B           U/L < TEST RNG  < 20.00   0.00076 R
40  2023-09-15 16:59:06 SMPL3   SAM           NH3B            mmol/L      1.817   0.05242 R
40  2023-09-15 16:59:16 SMPL3   SAM           LAC2B           g/L < TEST RNG  < 0.00    0.00329 R
40  2023-09-15 16:59:38 SMPL3   SAM           TP2D            g/L < TEST RNG  < 40.0    0.02702 R
40  2023-09-15 17:00:52 SMPL3   SAM           TP2LB           g/L     4.8   0.15436 R
40  2023-09-16 10:12:10 SMPL4   SAM           GLN2B           mmol/L      2.07    0.14503 R
40  2023-09-16 10:12:12 SMPL4   SAM           GLC3B           g/L     4.09    0.68160 R
40  2023-09-16 10:13:29 SMPL4   SAM           LDH2B           U/L     334.84    0.02665 R
40  2023-09-16 10:13:37 SMPL4   SAM           NH3B            mmol/L      3.788   0.11415 R
40  2023-09-16 10:22:55 SMPL4   SAM           LAC2B           g/L v   1.89    0.15187 R
```

Get familiar with the raw data
from this cell culture analyzer...



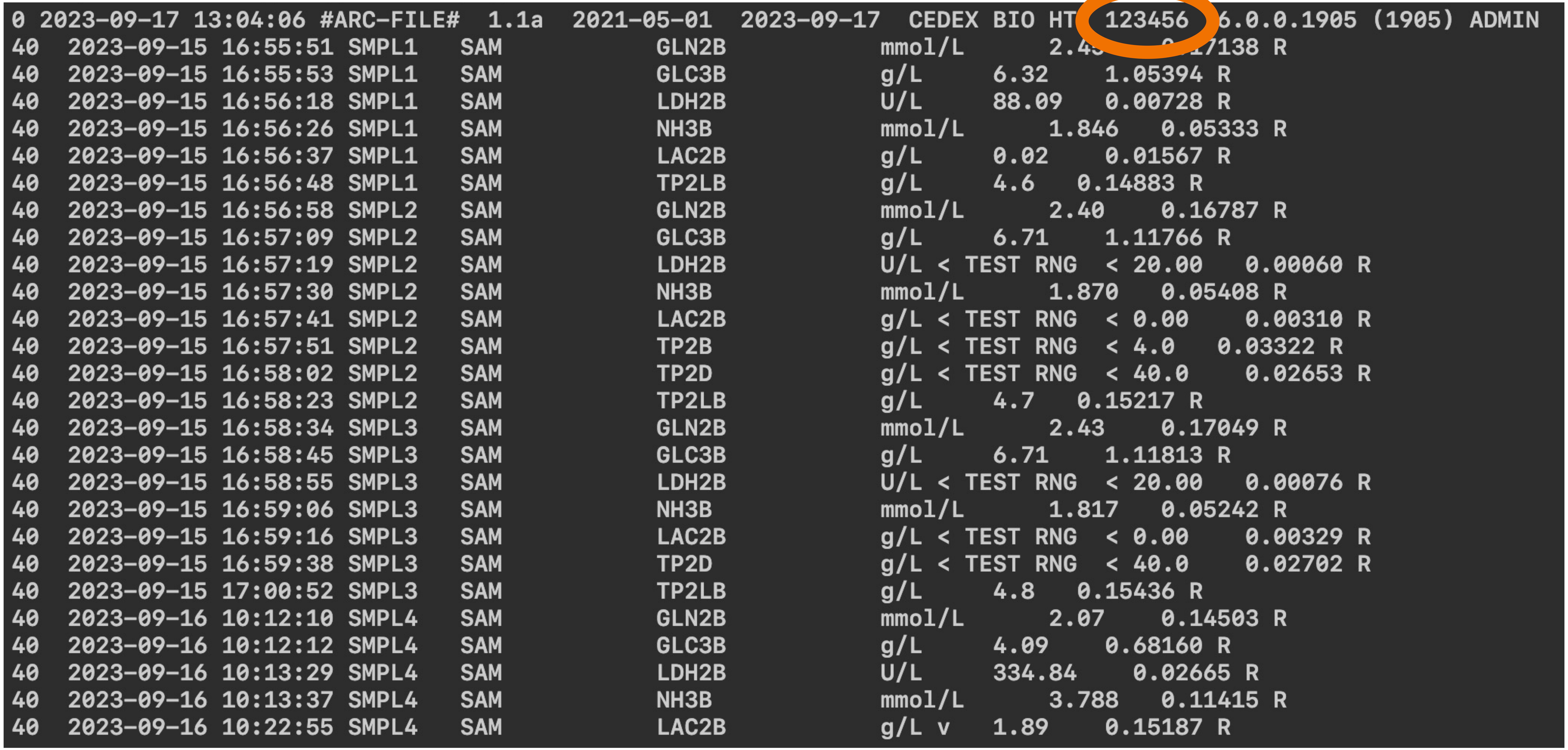
```
0 2023-09-17 13:04:06 #ARC-FILE#  1.1a  2021-05-01  2023-09-17  CEDEX BIO HT  123456  6.0.0.1905 (1905) ADMIN
40  2023-09-15 16:55:51 SMPL1   SAM           GLN2B           mmol/L      2.45    0.17138 R
40  2023-09-15 16:55:53 SMPL1   SAM           GLC3B           g/L     6.32    1.05394 R
40  2023-09-15 16:56:18 SMPL1   SAM           LDH2B           U/L     88.09   0.00728 R
40  2023-09-15 16:56:26 SMPL1   SAM           NH3B            mmol/L      1.846   0.05333 R
40  2023-09-15 16:56:37 SMPL1   SAM           LAC2B           g/L     0.02    0.01567 R
40  2023-09-15 16:56:48 SMPL1   SAM           TP2LB           g/L     4.6   0.14883 R
40  2023-09-15 16:56:58 SMPL2   SAM           GLN2B           mmol/L      2.40    0.16787 R
40  2023-09-15 16:57:09 SMPL2   SAM           GLC3B           g/L     6.71    1.11766 R
40  2023-09-15 16:57:19 SMPL2   SAM           LDH2B           U/L < TEST RNG < 20.00   0.00060 R
40  2023-09-15 16:57:30 SMPL2   SAM           NH3B            mmol/L      1.870   0.05408 R
40  2023-09-15 16:57:41 SMPL2   SAM           LAC2B           g/L < TEST RNG < 0.00    0.00310 R
40  2023-09-15 16:57:51 SMPL2   SAM           TP2B            g/L < TEST RNG < 4.0   0.03322 R
40  2023-09-15 16:58:02 SMPL2   SAM           TP2D            g/L < TEST RNG < 40.0    0.02653 R
40  2023-09-15 16:58:23 SMPL2   SAM           TP2LB           g/L     4.7   0.15217 R
40  2023-09-15 16:58:34 SMPL3   SAM           GLN2B           mmol/L      2.43    0.17049 R
40  2023-09-15 16:58:45 SMPL3   SAM           GLC3B           g/L     6.71    1.11813 R
40  2023-09-15 16:58:55 SMPL3   SAM           LDH2B           U/L < TEST RNG < 20.00   0.00076 R
40  2023-09-15 16:59:06 SMPL3   SAM           NH3B            mmol/L      1.817   0.05242 R
40  2023-09-15 16:59:16 SMPL3   SAM           LAC2B           g/L < TEST RNG < 0.00    0.00329 R
40  2023-09-15 16:59:38 SMPL3   SAM           TP2D            g/L < TEST RNG < 40.0    0.02702 R
40  2023-09-15 17:00:52 SMPL3   SAM           TP2LB           g/L     4.8   0.15436 R
40  2023-09-16 10:12:10 SMPL4   SAM           GLN2B           mmol/L      2.07    0.14503 R
40  2023-09-16 10:12:12 SMPL4   SAM           GLC3B           g/L     4.09    0.68160 R
40  2023-09-16 10:13:29 SMPL4   SAM           LDH2B           U/L     334.84    0.02665 R
40  2023-09-16 10:13:37 SMPL4   SAM           NH3B            mmol/L      3.788   0.11415 R
40  2023-09-16 10:22:55 SMPL4   SAM           LAC2B           g/L v   1.89    0.15187 R
```

Get familiar with the raw data
from this cell culture analyzer...



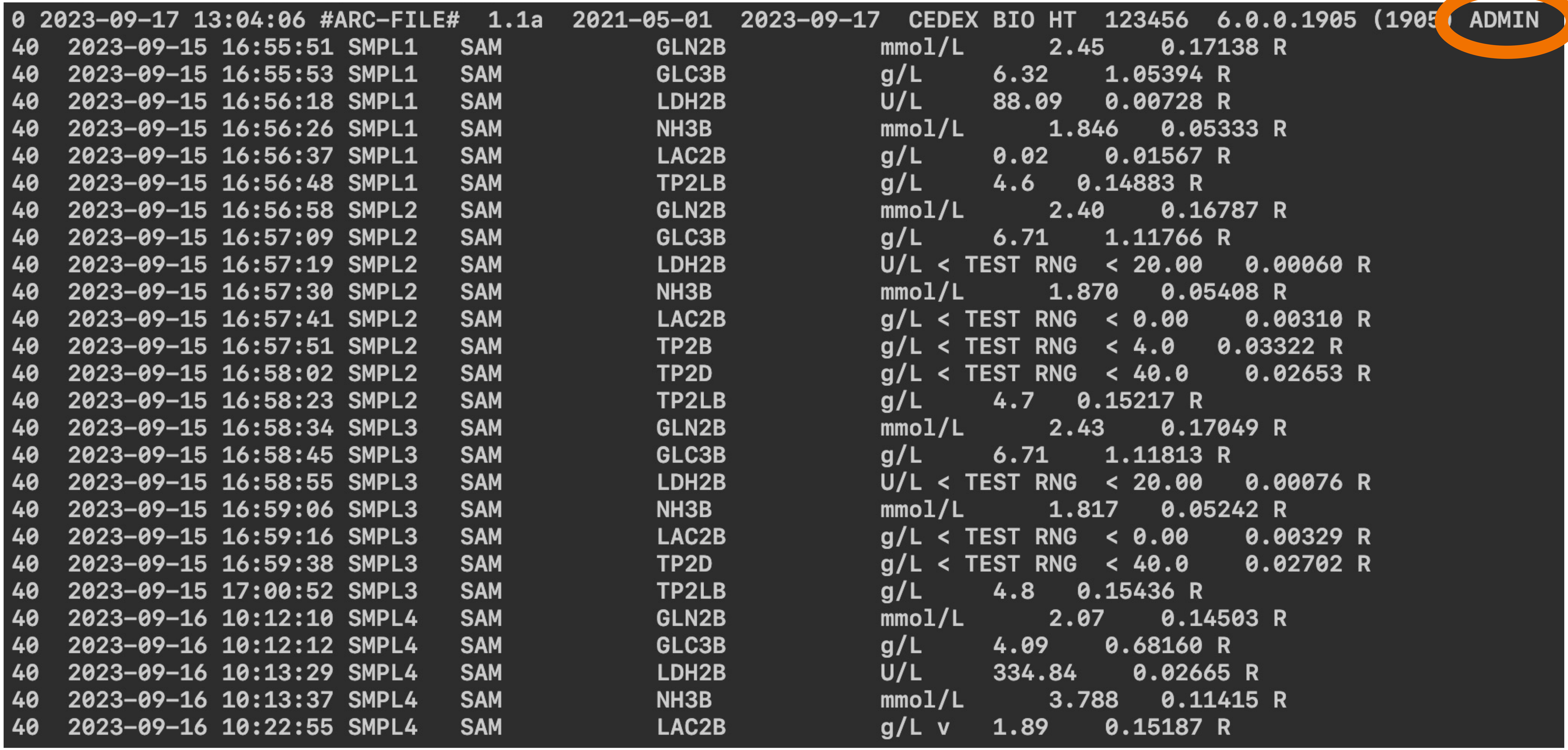
```
0 2023-09-17 13:04:06 #ARC-FILE#  1.1a  2021-05-01  2023-09-17  CEDEX BIO HT  123456  6.0.0.1905 (1905) ADMIN
40  2023-09-15 16:55:51 SMPL1   SAM           GLN2B           mmol/L      2.45    0.17138 R
40  2023-09-15 16:55:53 SMPL1   SAM           GLC3B           g/L     6.32    1.05394 R
40  2023-09-15 16:56:18 SMPL1   SAM           LDH2B           U/L     88.09   0.00728 R
40  2023-09-15 16:56:26 SMPL1   SAM           NH3B            mmol/L      1.846   0.05333 R
40  2023-09-15 16:56:37 SMPL1   SAM           LAC2B           g/L     0.02    0.01567 R
40  2023-09-15 16:56:48 SMPL1   SAM           TP2LB           g/L     4.6   0.14883 R
40  2023-09-15 16:56:58 SMPL2   SAM           GLN2B           mmol/L      2.40    0.16787 R
40  2023-09-15 16:57:09 SMPL2   SAM           GLC3B           g/L     6.71    1.11766 R
40  2023-09-15 16:57:19 SMPL2   SAM           LDH2B           U/L < TEST RNG  < 20.00   0.00060 R
40  2023-09-15 16:57:30 SMPL2   SAM           NH3B            mmol/L      1.870   0.05408 R
40  2023-09-15 16:57:41 SMPL2   SAM           LAC2B           g/L < TEST RNG  < 0.00    0.00310 R
40  2023-09-15 16:57:51 SMPL2   SAM           TP2B            g/L < TEST RNG  < 4.0   0.03322 R
40  2023-09-15 16:58:02 SMPL2   SAM           TP2D            g/L < TEST RNG  < 40.0    0.02653 R
40  2023-09-15 16:58:23 SMPL2   SAM           TP2LB           g/L     4.7   0.15217 R
40  2023-09-15 16:58:34 SMPL3   SAM           GLN2B           mmol/L      2.43    0.17049 R
40  2023-09-15 16:58:45 SMPL3   SAM           GLC3B           g/L     6.71    1.11813 R
40  2023-09-15 16:58:55 SMPL3   SAM           LDH2B           U/L < TEST RNG  < 20.00   0.00076 R
40  2023-09-15 16:59:06 SMPL3   SAM           NH3B            mmol/L      1.817   0.05242 R
40  2023-09-15 16:59:16 SMPL3   SAM           LAC2B           g/L < TEST RNG  < 0.00    0.00329 R
40  2023-09-15 16:59:38 SMPL3   SAM           TP2D            g/L < TEST RNG  < 40.0    0.02702 R
40  2023-09-15 17:00:52 SMPL3   SAM           TP2LB           g/L     4.8   0.15436 R
40  2023-09-16 10:12:10 SMPL4   SAM           GLN2B           mmol/L      2.07    0.14503 R
40  2023-09-16 10:12:12 SMPL4   SAM           GLC3B           g/L     4.09    0.68160 R
40  2023-09-16 10:13:29 SMPL4   SAM           LDH2B           U/L     334.84    0.02665 R
40  2023-09-16 10:13:37 SMPL4   SAM           NH3B            mmol/L      3.788   0.11415 R
40  2023-09-16 10:22:55 SMPL4   SAM           LAC2B           g/L v   1.89    0.15187 R
```

Get familiar with the raw data from this cell culture analyzer...



```
0 2023-09-17 13:04:06 #ARC-FILE#  1.1a  2021-05-01  2023-09-17  CEDEX BIO HT  123456  6.0.0.1905 (1905) ADMIN
40  2023-09-15 16:55:51 SMPL1   SAM           GLN2B           mmol/L      2.45    0.17138 R
40  2023-09-15 16:55:53 SMPL1   SAM           GLC3B           g/L     6.32    1.05394 R
40  2023-09-15 16:56:18 SMPL1   SAM           LDH2B           U/L     88.09   0.00728 R
40  2023-09-15 16:56:26 SMPL1   SAM           NH3B            mmol/L      1.846   0.05333 R
40  2023-09-15 16:56:37 SMPL1   SAM           LAC2B           g/L     0.02    0.01567 R
40  2023-09-15 16:56:48 SMPL1   SAM           TP2LB           g/L     4.6   0.14883 R
40  2023-09-15 16:56:58 SMPL2   SAM           GLN2B           mmol/L      2.40    0.16787 R
40  2023-09-15 16:57:09 SMPL2   SAM           GLC3B           g/L     6.71    1.11766 R
40  2023-09-15 16:57:19 SMPL2   SAM           LDH2B           U/L < TEST RNG  < 20.00   0.00060 R
40  2023-09-15 16:57:30 SMPL2   SAM           NH3B            mmol/L      1.870   0.05408 R
40  2023-09-15 16:57:41 SMPL2   SAM           LAC2B           g/L < TEST RNG  < 0.00    0.00310 R
40  2023-09-15 16:57:51 SMPL2   SAM           TP2B            g/L < TEST RNG  < 4.0   0.03322 R
40  2023-09-15 16:58:02 SMPL2   SAM           TP2D            g/L < TEST RNG  < 40.0    0.02653 R
40  2023-09-15 16:58:23 SMPL2   SAM           TP2LB           g/L     4.7   0.15217 R
40  2023-09-15 16:58:34 SMPL3   SAM           GLN2B           mmol/L      2.43    0.17049 R
40  2023-09-15 16:58:45 SMPL3   SAM           GLC3B           g/L     6.71    1.11813 R
40  2023-09-15 16:58:55 SMPL3   SAM           LDH2B           U/L < TEST RNG  < 20.00   0.00076 R
40  2023-09-15 16:59:06 SMPL3   SAM           NH3B            mmol/L      1.817   0.05242 R
40  2023-09-15 16:59:16 SMPL3   SAM           LAC2B           g/L < TEST RNG  < 0.00    0.00329 R
40  2023-09-15 16:59:38 SMPL3   SAM           TP2D            g/L < TEST RNG  < 40.0    0.02702 R
40  2023-09-15 17:00:52 SMPL3   SAM           TP2LB           g/L     4.8   0.15436 R
40  2023-09-16 10:12:10 SMPL4   SAM           GLN2B           mmol/L      2.07    0.14503 R
40  2023-09-16 10:12:12 SMPL4   SAM           GLC3B           g/L     4.09    0.68160 R
40  2023-09-16 10:13:29 SMPL4   SAM           LDH2B           U/L     334.84    0.02665 R
40  2023-09-16 10:13:37 SMPL4   SAM           NH3B            mmol/L      3.788   0.11415 R
40  2023-09-16 10:22:55 SMPL4   SAM           LAC2B           g/L v   1.89    0.15187 R
```

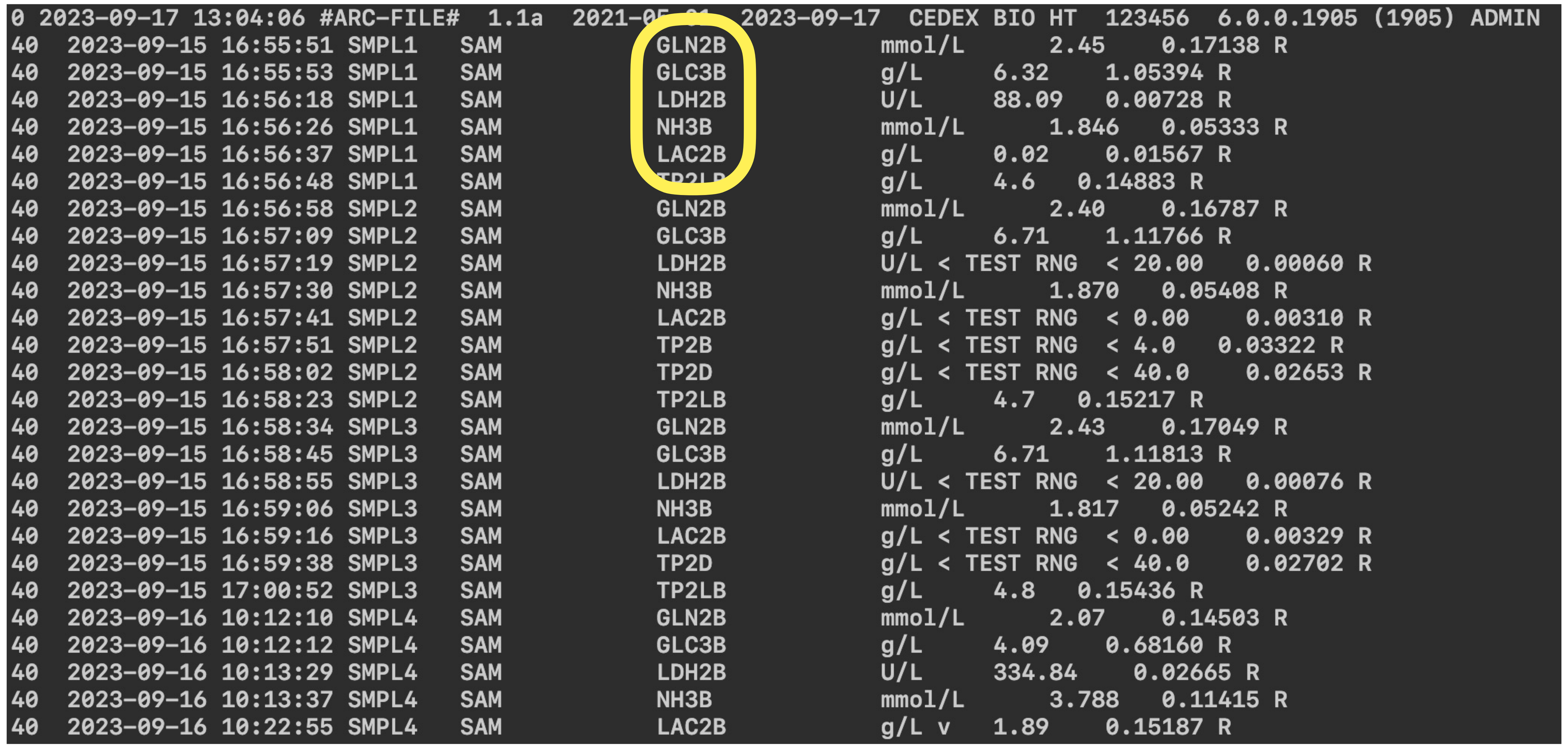
Get familiar with the raw data from this cell culture analyzer...



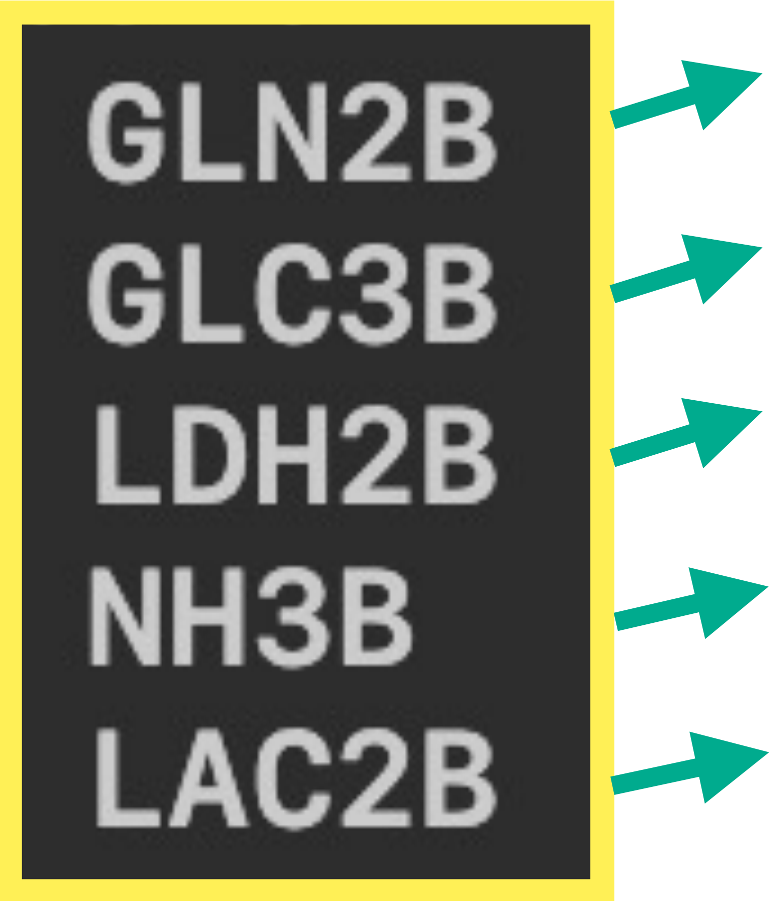
```
0 2023-09-17 13:04:06 #ARC-FILE#  1.1a  2021-05-01  2023-09-17  CEDEX BIO HT  123456  6.0.0.1905 (1905  ADMIN
40  2023-09-15 16:55:51 SMPL1   SAM           GLN2B           mmol/L      2.45    0.17138 R
40  2023-09-15 16:55:53 SMPL1   SAM           GLC3B           g/L     6.32    1.05394 R
40  2023-09-15 16:56:18 SMPL1   SAM           LDH2B           U/L     88.09   0.00728 R
40  2023-09-15 16:56:26 SMPL1   SAM           NH3B            mmol/L      1.846   0.05333 R
40  2023-09-15 16:56:37 SMPL1   SAM           LAC2B           g/L     0.02    0.01567 R
40  2023-09-15 16:56:48 SMPL1   SAM           TP2LB           g/L     4.6   0.14883 R
40  2023-09-15 16:56:58 SMPL2   SAM           GLN2B           mmol/L      2.40    0.16787 R
40  2023-09-15 16:57:09 SMPL2   SAM           GLC3B           g/L     6.71    1.11766 R
40  2023-09-15 16:57:19 SMPL2   SAM           LDH2B           U/L < TEST RNG  < 20.00   0.00060 R
40  2023-09-15 16:57:30 SMPL2   SAM           NH3B            mmol/L      1.870   0.05408 R
40  2023-09-15 16:57:41 SMPL2   SAM           LAC2B           g/L < TEST RNG  < 0.00    0.00310 R
40  2023-09-15 16:57:51 SMPL2   SAM           TP2B            g/L < TEST RNG  < 4.0   0.03322 R
40  2023-09-15 16:58:02 SMPL2   SAM           TP2D            g/L < TEST RNG  < 40.0    0.02653 R
40  2023-09-15 16:58:23 SMPL2   SAM           TP2LB           g/L     4.7   0.15217 R
40  2023-09-15 16:58:34 SMPL3   SAM           GLN2B           mmol/L      2.43    0.17049 R
40  2023-09-15 16:58:45 SMPL3   SAM           GLC3B           g/L     6.71    1.11813 R
40  2023-09-15 16:58:55 SMPL3   SAM           LDH2B           U/L < TEST RNG  < 20.00   0.00076 R
40  2023-09-15 16:59:06 SMPL3   SAM           NH3B            mmol/L      1.817   0.05242 R
40  2023-09-15 16:59:16 SMPL3   SAM           LAC2B           g/L < TEST RNG  < 0.00    0.00329 R
40  2023-09-15 16:59:38 SMPL3   SAM           TP2D            g/L < TEST RNG  < 40.0    0.02702 R
40  2023-09-15 17:00:52 SMPL3   SAM           TP2LB           g/L     4.8   0.15436 R
40  2023-09-16 10:12:10 SMPL4   SAM           GLN2B           mmol/L      2.07    0.14503 R
40  2023-09-16 10:12:12 SMPL4   SAM           GLC3B           g/L     4.09    0.68160 R
40  2023-09-16 10:13:29 SMPL4   SAM           LDH2B           U/L     334.84    0.02665 R
40  2023-09-16 10:13:37 SMPL4   SAM           NH3B            mmol/L      3.788   0.11415 R
40  2023-09-16 10:22:55 SMPL4   SAM           LAC2B           g/L v   1.89    0.15187 R
```

Our findings: what is an LLM good at?

Domain knowledge

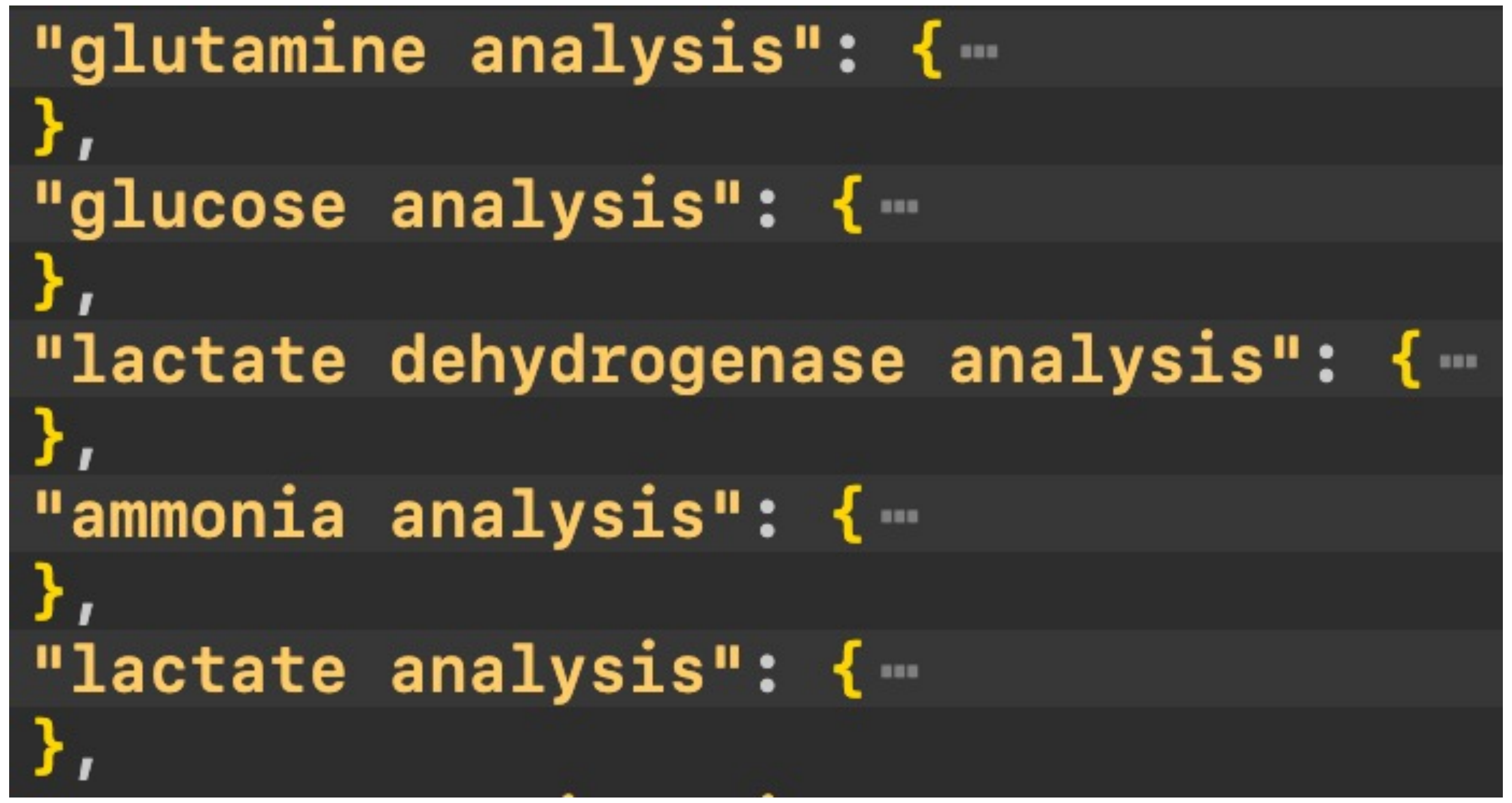


```
0 2023-09-17 13:04:06 #ARC-FILE#  1.1a  2021-05-01  2023-09-17  CEDEX BIO HT  123456  6.0.0.1905 (1905) ADMIN
40  2023-09-15 16:55:51 SMPL1    SAM            GLN2B            mmol/L      2.45    0.17138 R
40  2023-09-15 16:55:53 SMPL1    SAM            GLC3B            g/L     6.32    1.05394 R
40  2023-09-15 16:56:18 SMPL1    SAM            LDH2B            U/L     88.09   0.00728 R
40  2023-09-15 16:56:26 SMPL1    SAM            NH3B             mmol/L      1.846   0.05333 R
40  2023-09-15 16:56:37 SMPL1    SAM            LAC2B            g/L     0.02    0.01567 R
40  2023-09-15 16:56:48 SMPL1    SAM            TP2LB            g/L     4.6   0.14883 R
40  2023-09-15 16:56:58 SMPL2    SAM            GLN2B            mmol/L      2.40    0.16787 R
40  2023-09-15 16:57:09 SMPL2    SAM            GLC3B            g/L     6.71    1.11766 R
40  2023-09-15 16:57:19 SMPL2    SAM            LDH2B            U/L < TEST RNG  < 20.00   0.00060 R
40  2023-09-15 16:57:30 SMPL2    SAM            NH3B             mmol/L      1.870   0.05408 R
40  2023-09-15 16:57:41 SMPL2    SAM            LAC2B            g/L < TEST RNG  < 0.00    0.00310 R
40  2023-09-15 16:57:51 SMPL2    SAM            TP2B             g/L < TEST RNG  < 4.0   0.03322 R
40  2023-09-15 16:58:02 SMPL2    SAM            TP2D             g/L < TEST RNG  < 40.0    0.02653 R
40  2023-09-15 16:58:23 SMPL2    SAM            TP2LB            g/L     4.7   0.15217 R
40  2023-09-15 16:58:34 SMPL3    SAM            GLN2B            mmol/L      2.43    0.17049 R
40  2023-09-15 16:58:45 SMPL3    SAM            GLC3B            g/L     6.71    1.11813 R
40  2023-09-15 16:58:55 SMPL3    SAM            LDH2B            U/L < TEST RNG  < 20.00   0.00076 R
40  2023-09-15 16:59:06 SMPL3    SAM            NH3B             mmol/L      1.817   0.05242 R
40  2023-09-15 16:59:16 SMPL3    SAM            LAC2B            g/L < TEST RNG  < 0.00    0.00329 R
40  2023-09-15 16:59:38 SMPL3    SAM            TP2D             g/L < TEST RNG  < 40.0    0.02702 R
40  2023-09-15 17:00:52 SMPL3    SAM            TP2LB            g/L     4.8   0.15436 R
40  2023-09-16 10:12:10 SMPL4    SAM            GLN2B            mmol/L      2.07    0.14503 R
40  2023-09-16 10:12:12 SMPL4    SAM            GLC3B            g/L     4.09    0.68160 R
40  2023-09-16 10:13:29 SMPL4    SAM            LDH2B            U/L     334.84    0.02665 R
40  2023-09-16 10:13:37 SMPL4    SAM            NH3B             mmol/L      3.788   0.11415 R
40  2023-09-16 10:22:55 SMPL4    SAM            LAC2B            g/L v   1.89    0.15187 R
```



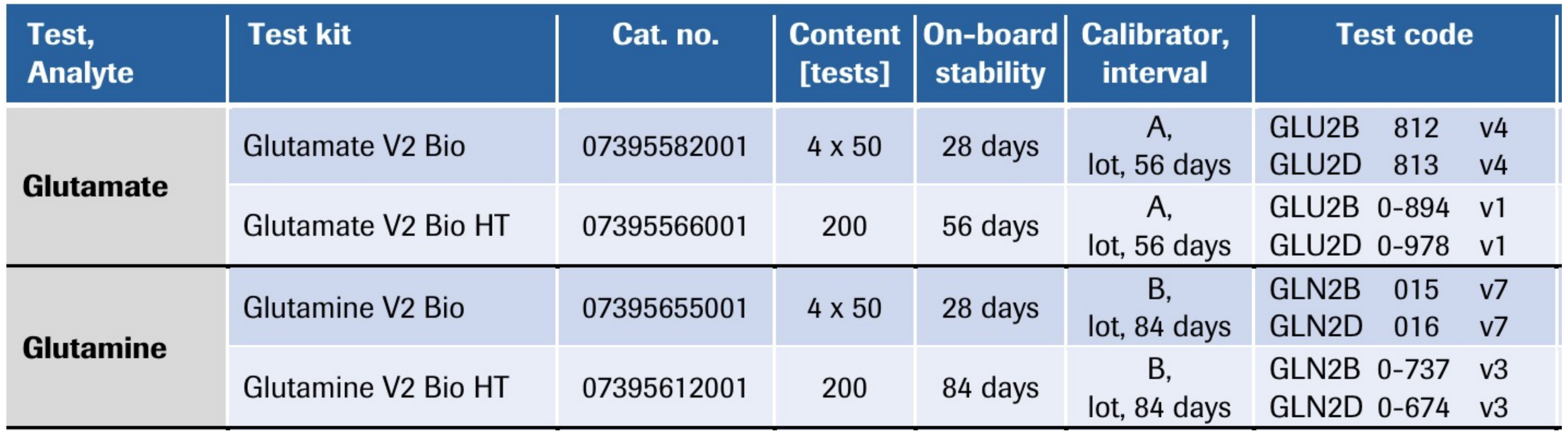
```
GLN2B
GLC3B
LDH2B
NH3B
LAC2B
```

Raw instrument data



```
"glutamine analysis": {⋯
},
"glucose analysis": {⋯
},
"lactate dehydrogenase analysis": {⋯
},
"ammonia analysis": {⋯
},
"lactate analysis": {⋯
},
```

An LLM extraction result (collapsed for readability)

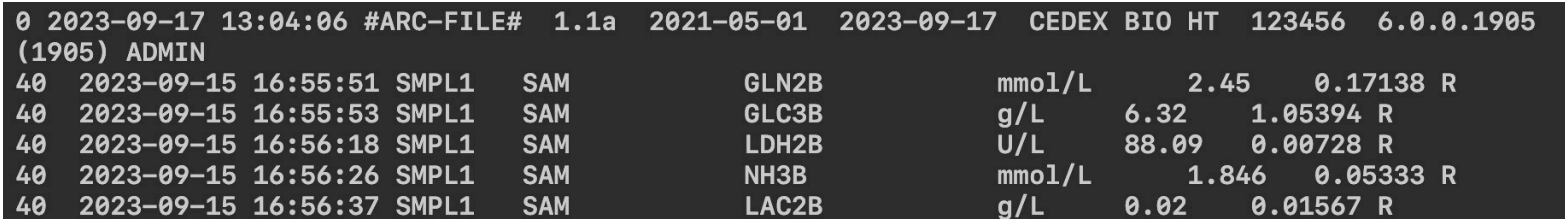


Test, Analyte	Test kit	Cat. no.	Content [tests]	On-board stability	Calibrator, interval	Test code
Glutamate	Glutamate V2 Bio	07395582001	4 x 50	28 days	A, lot, 56 days	GLU2B 812 v4 GLU2D 813 v4
	Glutamate V2 Bio HT	07395566001	200	56 days	A, lot, 56 days	GLU2B 0-894 v1 GLU2D 0-978 v1
Glutamine	Glutamine V2 Bio	07395655001	4 x 50	28 days	B, lot, 84 days	GLN2B 015 v7 GLN2D 016 v7
	Glutamine V2 Bio HT	07395612001	200	84 days	B, lot, 84 days	GLN2B 0-737 v3 GLN2D 0-674 v3

Unlike an engineer, doesn't need back-and-forth with SMEs for domain knowledge

Our findings: what is an LLM good at?

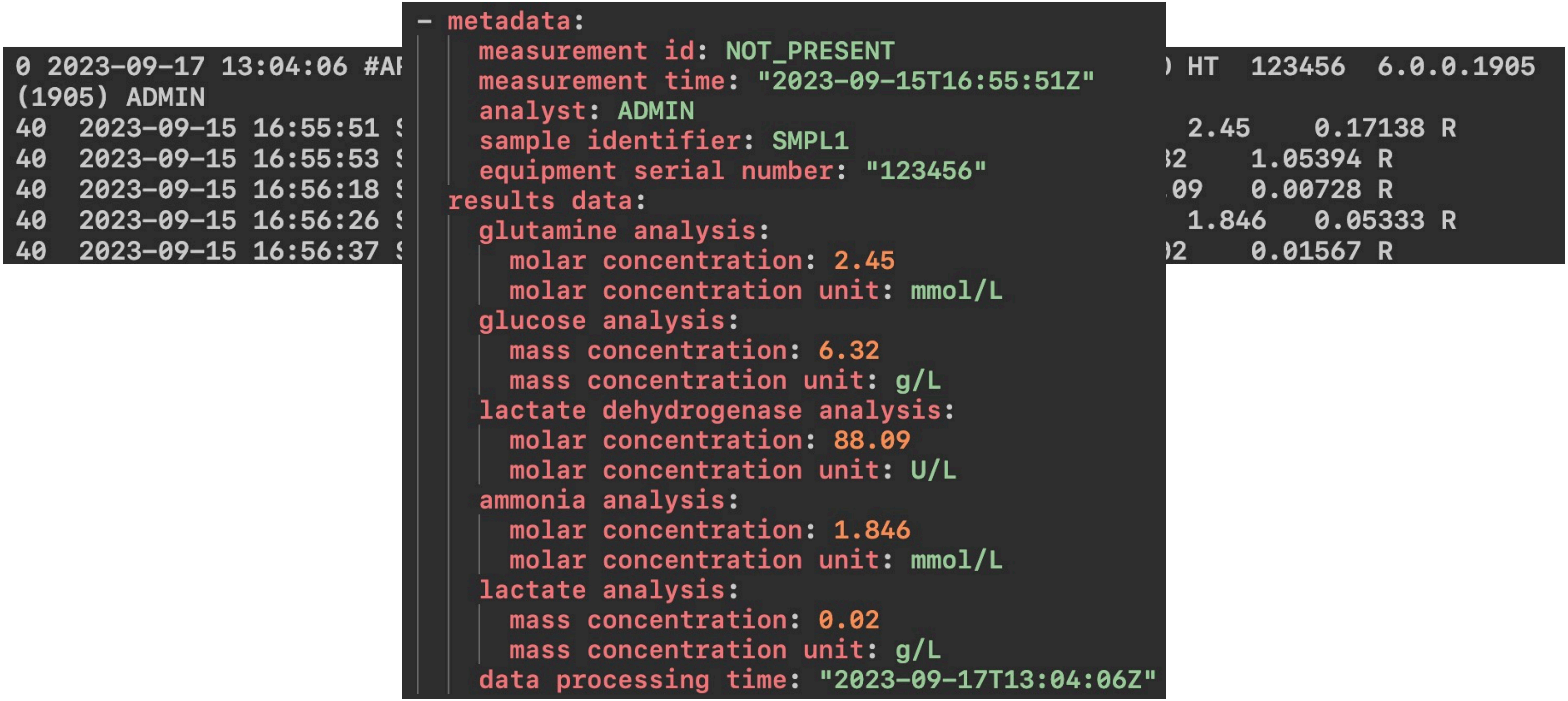
Extracting and Structuring Data



```
0 2023-09-17 13:04:06 #ARC-FILE#  1.1a  2021-05-01  2023-09-17  CEDEX BIO HT  123456  6.0.0.1905
(1905) ADMIN
40  2023-09-15 16:55:51 SMPL1   SAM           GLN2B           mmol/L      2.45    0.17138 R
40  2023-09-15 16:55:53 SMPL1   SAM           GLC3B           g/L     6.32    1.05394 R
40  2023-09-15 16:56:18 SMPL1   SAM           LDH2B           U/L     88.09   0.00728 R
40  2023-09-15 16:56:26 SMPL1   SAM           NH3B            mmol/L      1.846   0.05333 R
40  2023-09-15 16:56:37 SMPL1   SAM           LAC2B           g/L     0.02    0.01567 R
```

Our findings: what is an LLM good at?

Extracting and Structuring Data



```
0 2023-09-17 13:04:06 #AI
(1905) ADMIN
40  2023-09-15 16:55:51 S
40  2023-09-15 16:55:53 S
40  2023-09-15 16:56:18 S
40  2023-09-15 16:56:26 S
40  2023-09-15 16:56:37 S
```

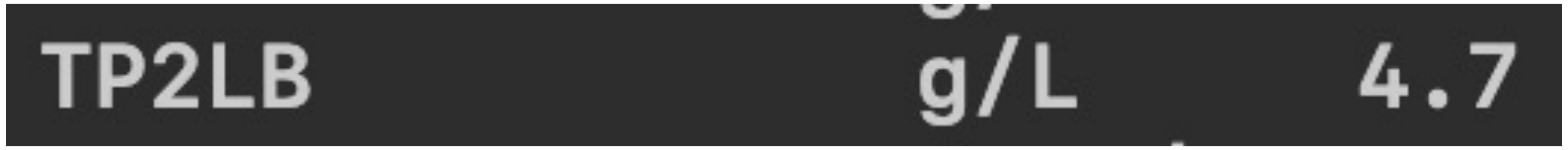
```
) HT  123456  6.0.0.1905

 2.45     0.17138 R
32     1.05394 R
.09    0.00728 R
 1.846    0.05333 R
)2     0.01567 R
```

```
- metadata:
    measurement id: NOT_PRESENT
    measurement time: "2023-09-15T16:55:51Z"
    analyst: ADMIN
    sample identifier: SMPL1
    equipment serial number: "123456"
  results data:
    glutamine analysis:
      molar concentration: 2.45
      molar concentration unit: mmol/L
    glucose analysis:
      mass concentration: 6.32
      mass concentration unit: g/L
    lactate dehydrogenase analysis:
      molar concentration: 88.09
      molar concentration unit: U/L
    ammonia analysis:
      molar concentration: 1.846
      molar concentration unit: mmol/L
    lactate analysis:
      mass concentration: 0.02
      mass concentration unit: g/L
    data processing time: "2023-09-17T13:04:06Z"
```

Our findings: what is an LLM good at?

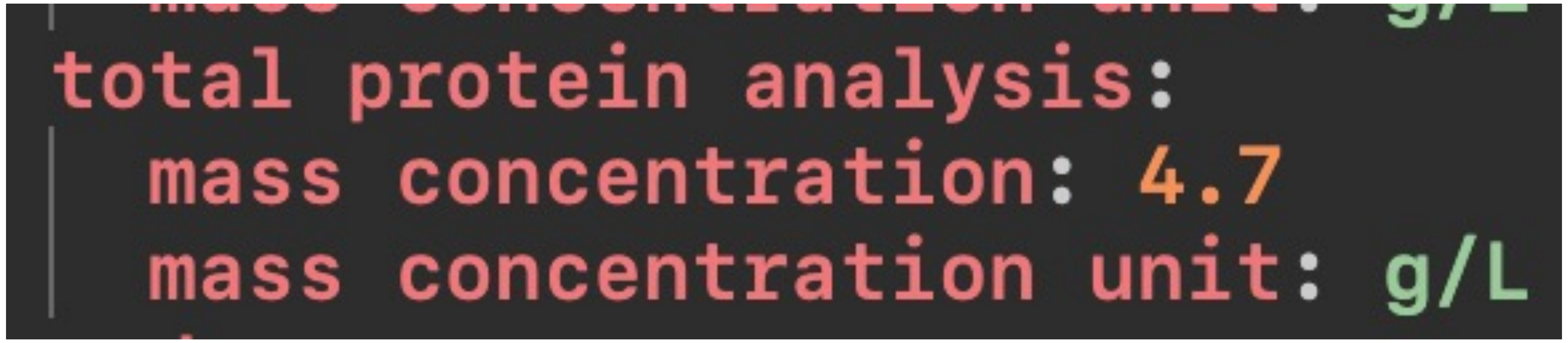
Generalization

Raw instrument data



```
TP2LB                          g/L         4.7
```

Successful extraction of field not in ASM schema

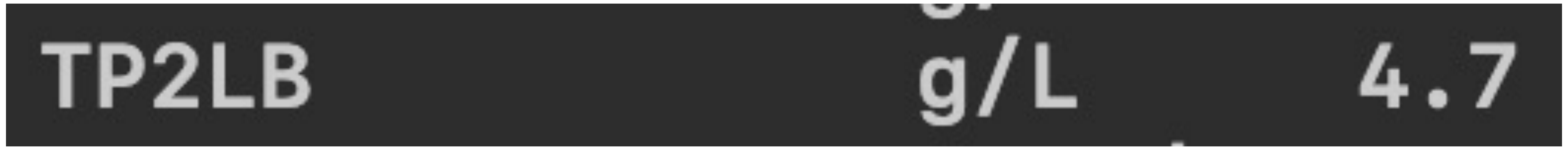


```
total protein analysis:
  mass concentration: 4.7
  mass concentration unit: g/L
```

Our findings: what is an LLM good at?

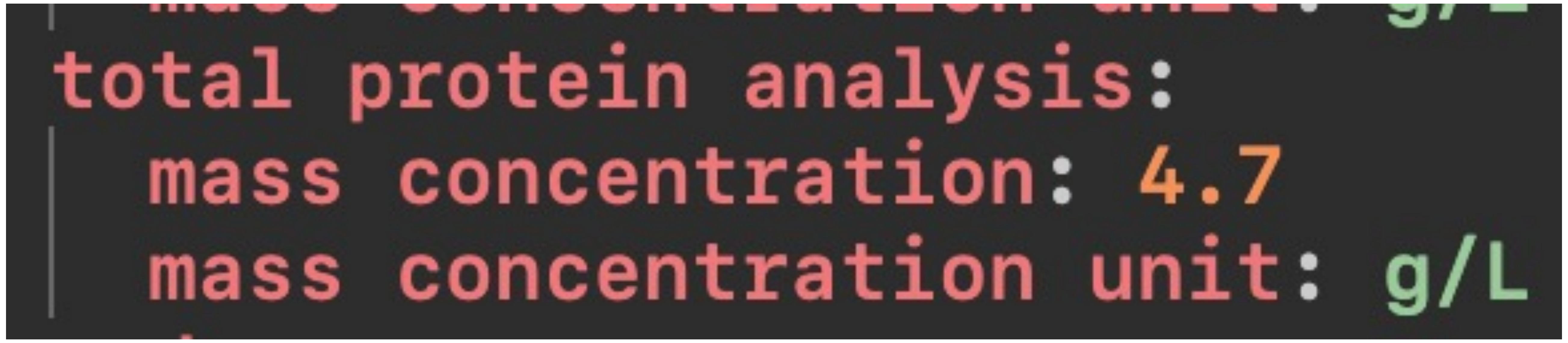
Generalization

Raw instrument data



```
TP2LB                     g/L          4.7
```

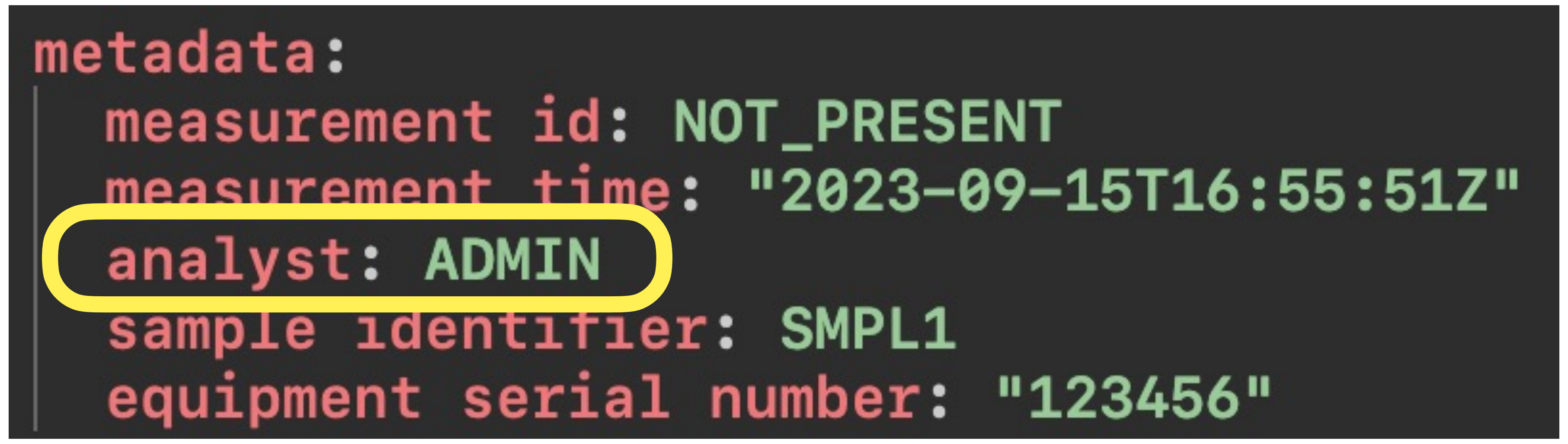
Successful extraction of field not in ASM schema



```
total protein analysis:
  mass concentration: 4.7
  mass concentration unit: g/L
```

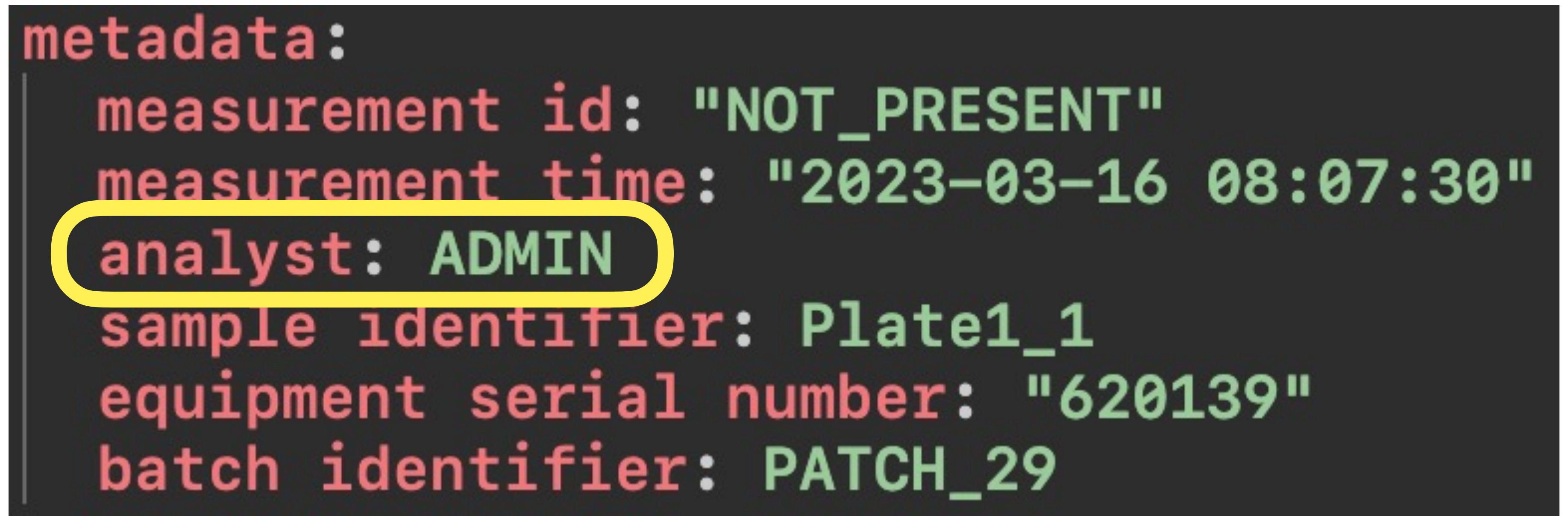
Our findings: what is an LLM bad at?

Consistency



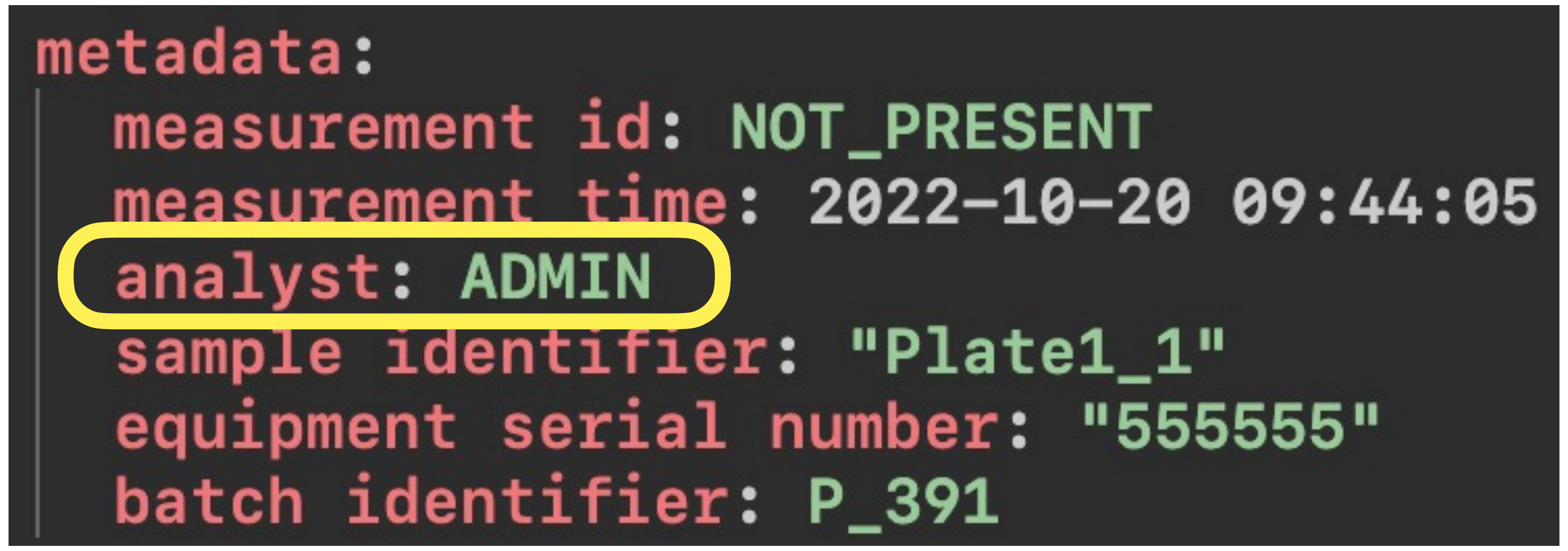
```
metadata:
  measurement id: NOT_PRESENT
  measurement time: "2023-09-15T16:55:51Z"
  analyst: ADMIN
  sample identifier: SMPL1
  equipment serial number: "123456"
```

✅



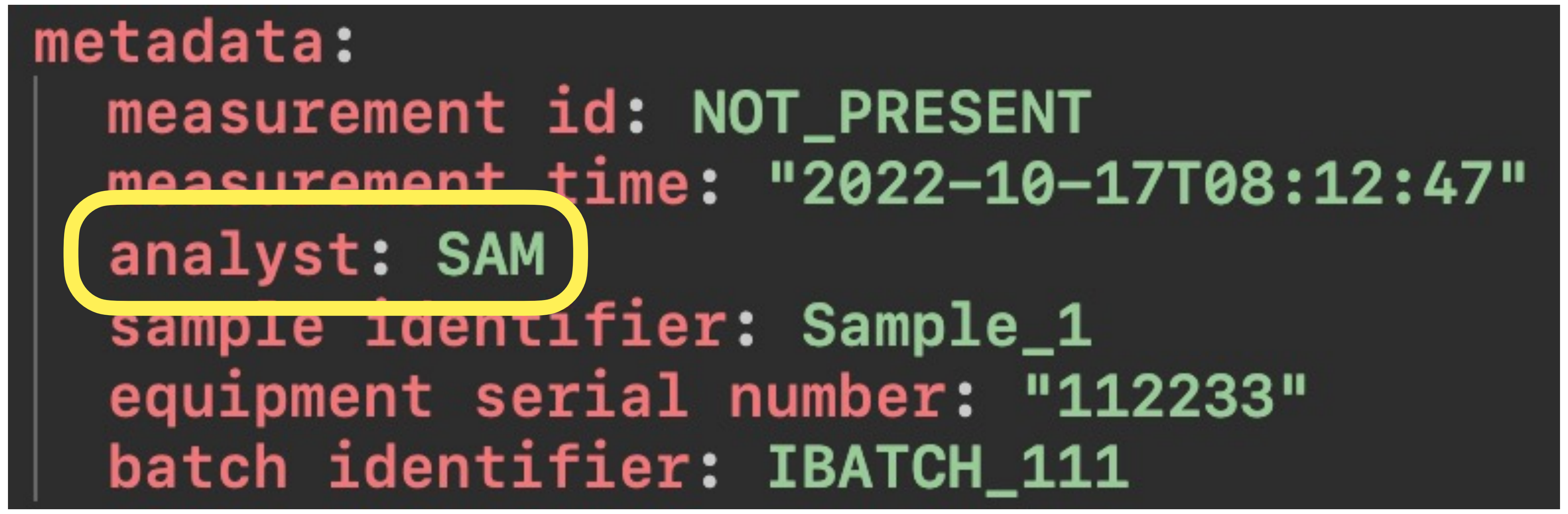
```
metadata:
  measurement id: "NOT_PRESENT"
  measurement time: "2023-03-16 08:07:30"
  analyst: ADMIN
  sample identifier: Plate1_1
  equipment serial number: "620139"
  batch identifier: PATCH_29
```

✅



```
metadata:
  measurement id: NOT_PRESENT
  measurement time: 2022-10-20 09:44:05
  analyst: ADMIN
  sample identifier: "Plate1_1"
  equipment serial number: "555555"
  batch identifier: P_391
```

✅



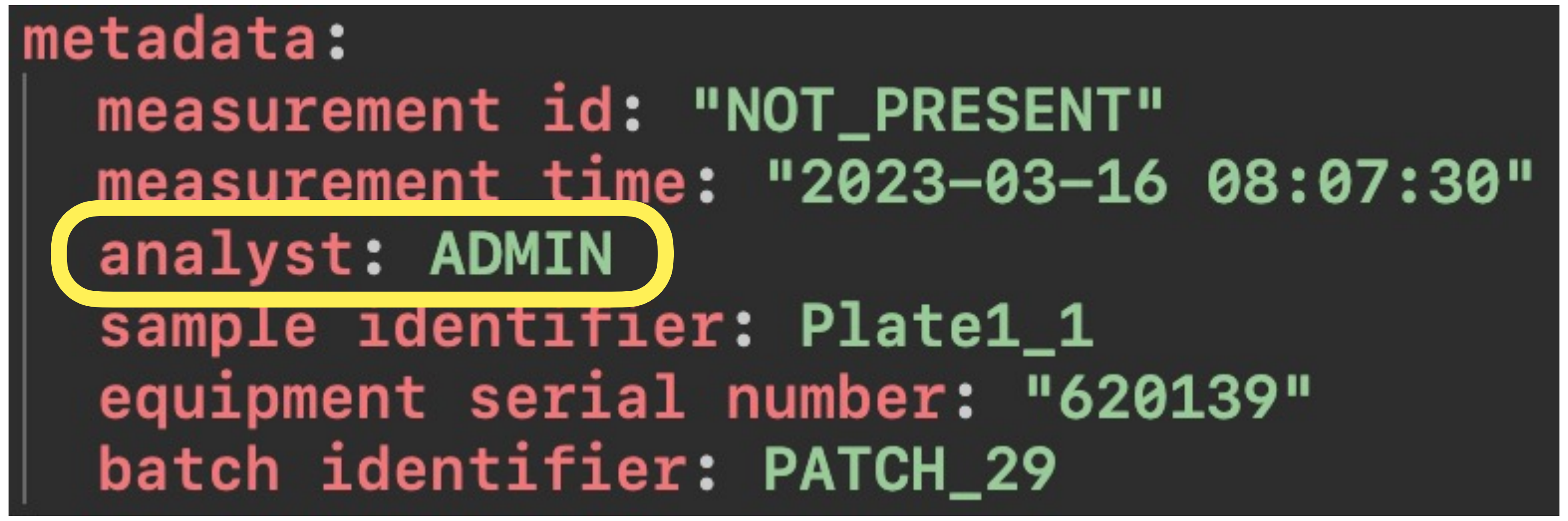
```
metadata:
  measurement id: NOT_PRESENT
  measurement time: "2022-10-17T08:12:47"
  analyst: SAM
  sample identifier: Sample_1
  equipment serial number: "112233"
  batch identifier: IBATCH_111
```

❌

Needs validation!

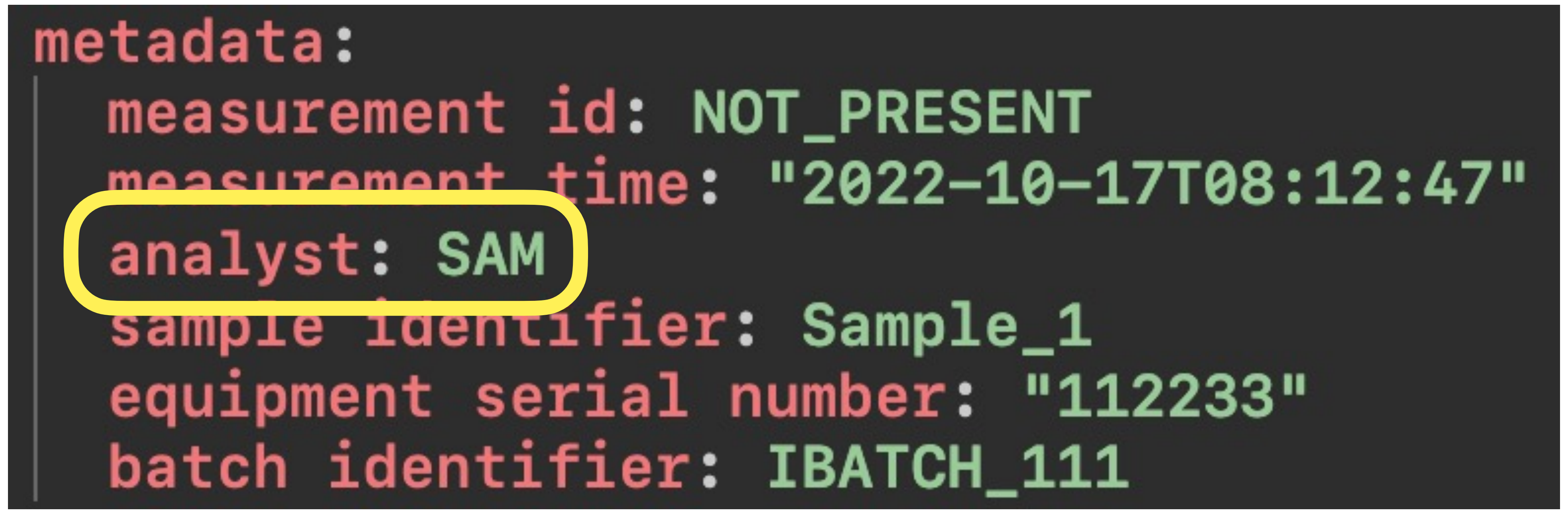
Our findings: what is an LLM bad at?

Explainability



```
metadata:
  measurement id: "NOT_PRESENT"
  measurement time: "2023-03-16 08:07:30"
  analyst: ADMIN
  sample identifier: Plate1_1
  equipment serial number: "620139"
  batch identifier: PATCH_29
```

✅



```
metadata:
  measurement id: NOT_PRESENT
  measurement time: "2022-10-17T08:12:47"
  analyst: SAM
  sample identifier: Sample_1
  equipment serial number: "112233"
  batch identifier: IBATCH_111
```

❌

How do you explain this to an auditor?

Our findings: what is an LLM bad at?

Expensive to run at high volumes

Moderately-sized raw data file (~50k tokens)	**~$5 per day**
If you’re a lab with 10k instruments	**10s of millions annually ($)**

Many caveats — just for illustration

Our findings: what is an LLM bad at?

Expensive to run at high volumes

Moderately-sized raw data file (~50k tokens)	**~$5 per day**
If you’re a lab with 10k instruments	**10s of millions annually ($)**

💡 All three LLM negatives (**consistency**, **explainability**, **cost**) can be solved through code

Our approach

Take advantage of the good and mitigate the bad

PYTHON

converter.py

Our approach

Take advantage of the good and mitigate the bad

raw instrument data

raw instrument data

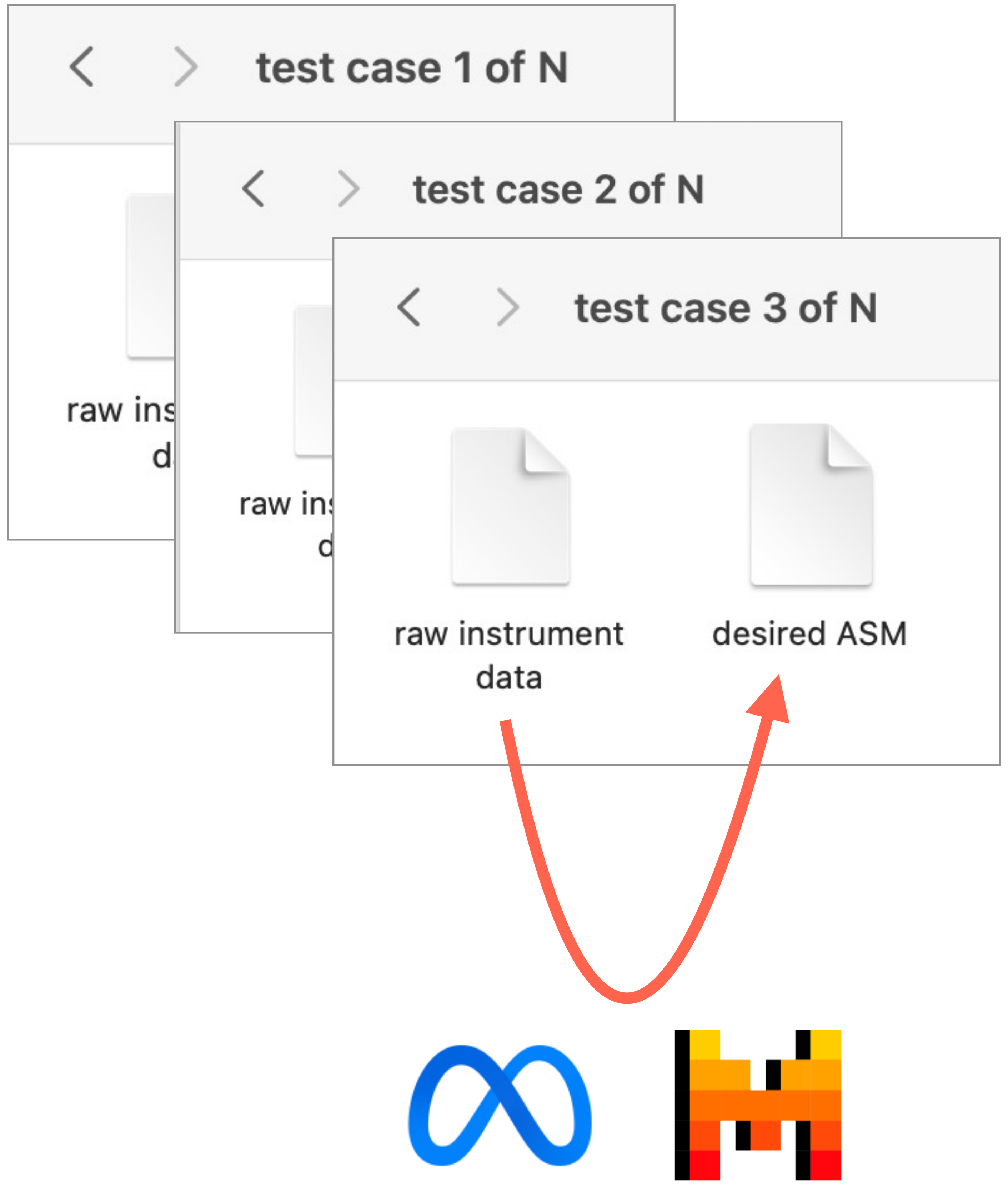
raw instrument data

PYTHON

converter.py

Our approach

Take advantage of the good and mitigate the bad



PYTHON

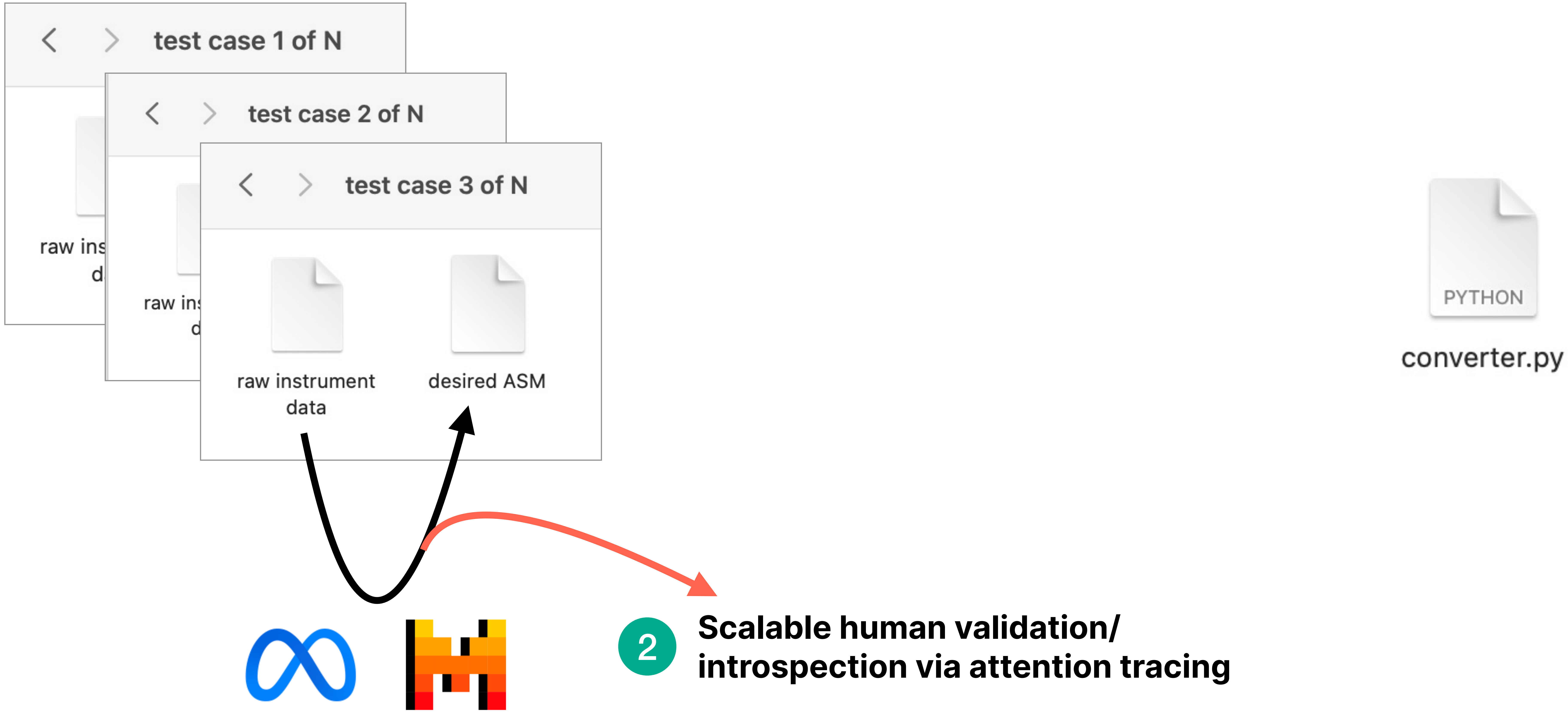
converter.py

1 Test set generation with LLMs

*open-weights required
for attention tracing

Our approach

Take advantage of the good and mitigate the bad



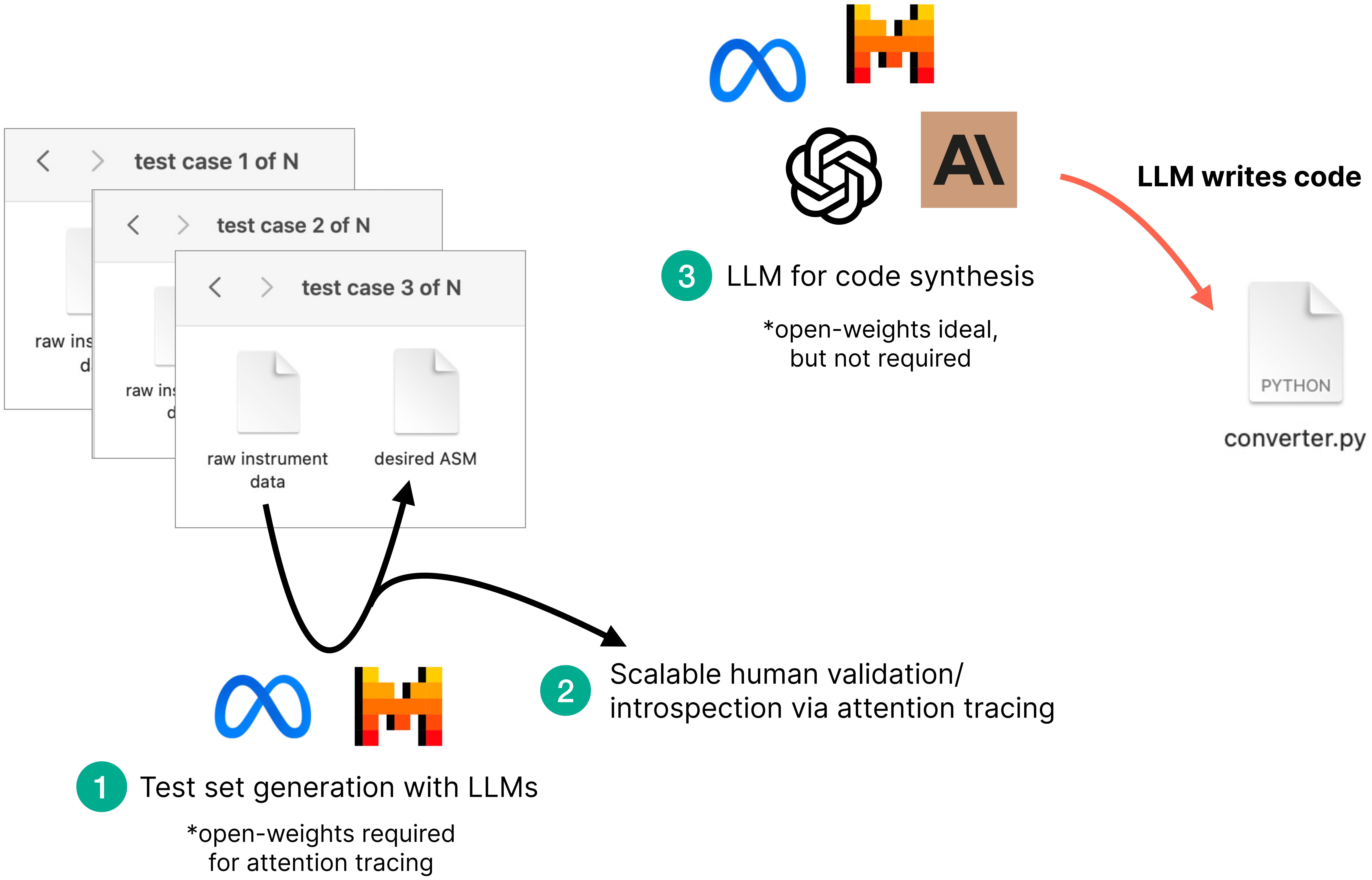
2 **Scalable human validation/ introspection via attention tracing**

1 Test set generation with LLMs

*open-weights required
for attention tracing

Our approach

Take advantage of the good and mitigate the bad



1 Test set generation with LLMs

*open-weights required for attention tracing

2 Scalable human validation/ introspection via attention tracing

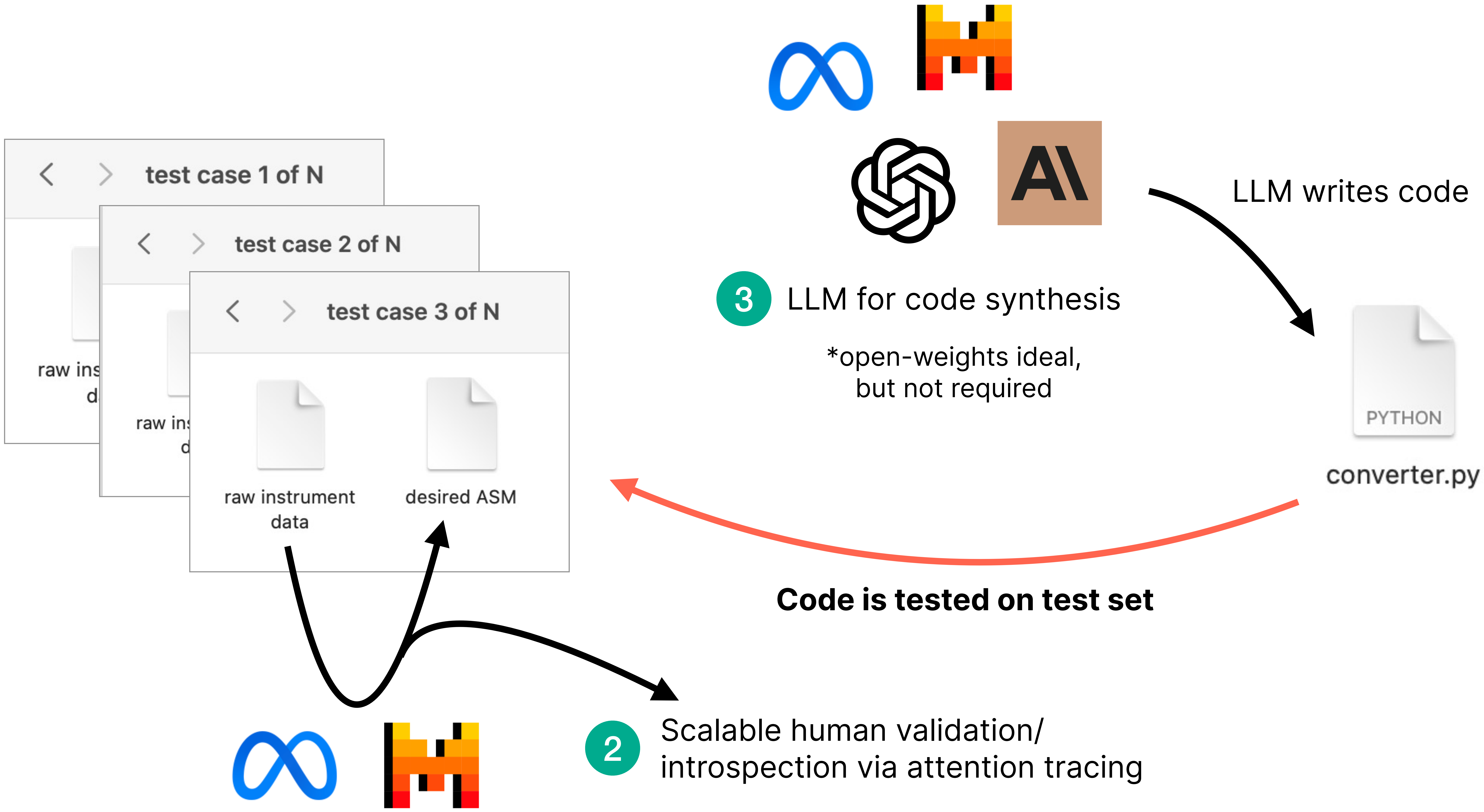
3 LLM for code synthesis

*open-weights ideal, but not required

LLM writes code

Our approach

Take advantage of the good and mitigate the bad



LLM writes code

3 LLM for code synthesis

*open-weights ideal, but not required

Code is tested on test set

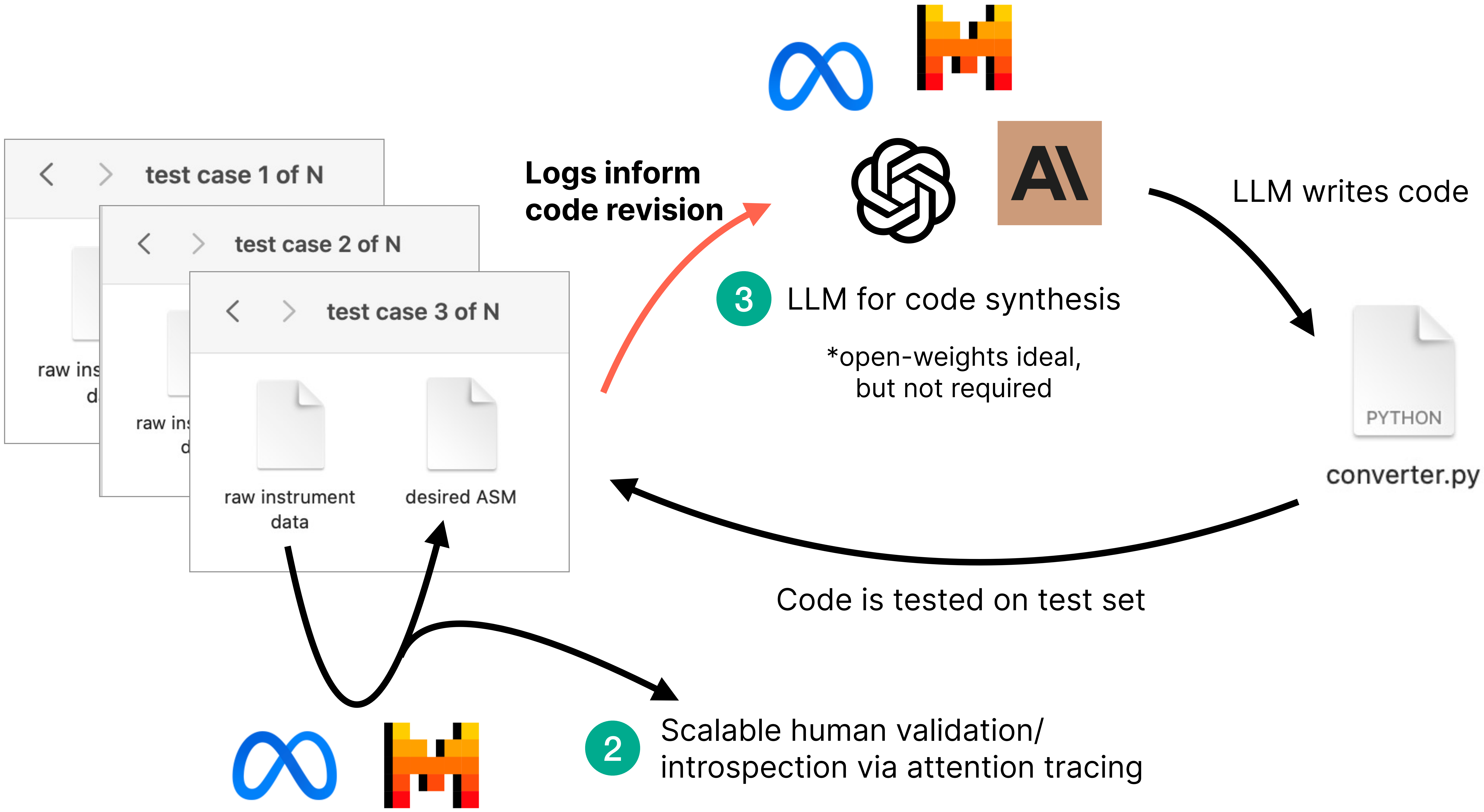
2 Scalable human validation/ introspection via attention tracing

1 Test set generation with LLMs

*open-weights required for attention tracing

Our approach

Take advantage of the good and mitigate the bad



Logs inform code revision

3 LLM for code synthesis

*open-weights ideal, but not required

LLM writes code

Code is tested on test set

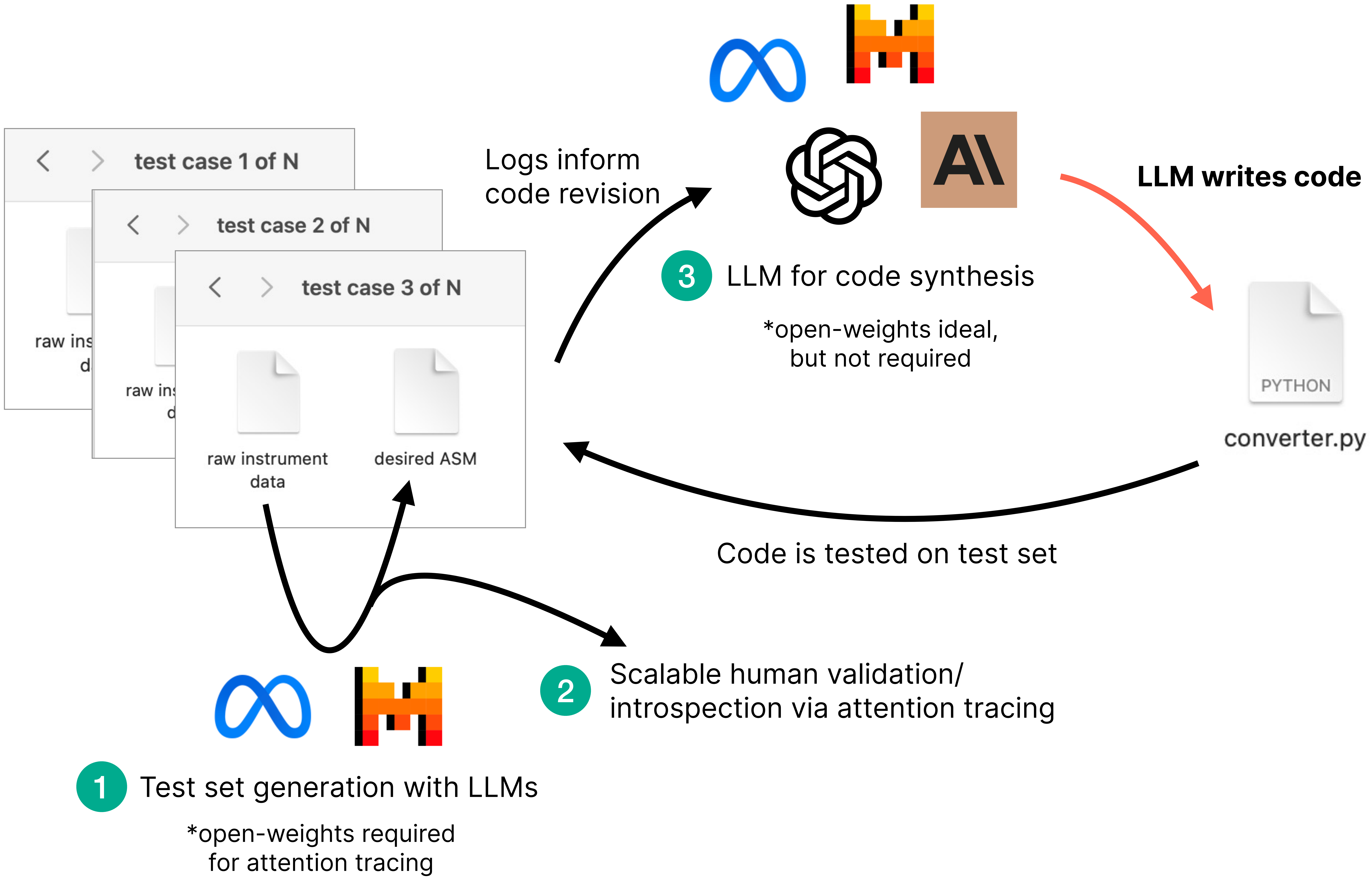
2 Scalable human validation/ introspection via attention tracing

1 Test set generation with LLMs

*open-weights required for attention tracing

Our approach

Take advantage of the good and mitigate the bad



test case 1 of N

raw ins d

test case 2 of N

raw ins d

test case 3 of N

raw instrument data

desired ASM

Logs inform code revision

3 LLM for code synthesis

*open-weights ideal, but not required

LLM writes code

PYTHON

converter.py

Code is tested on test set

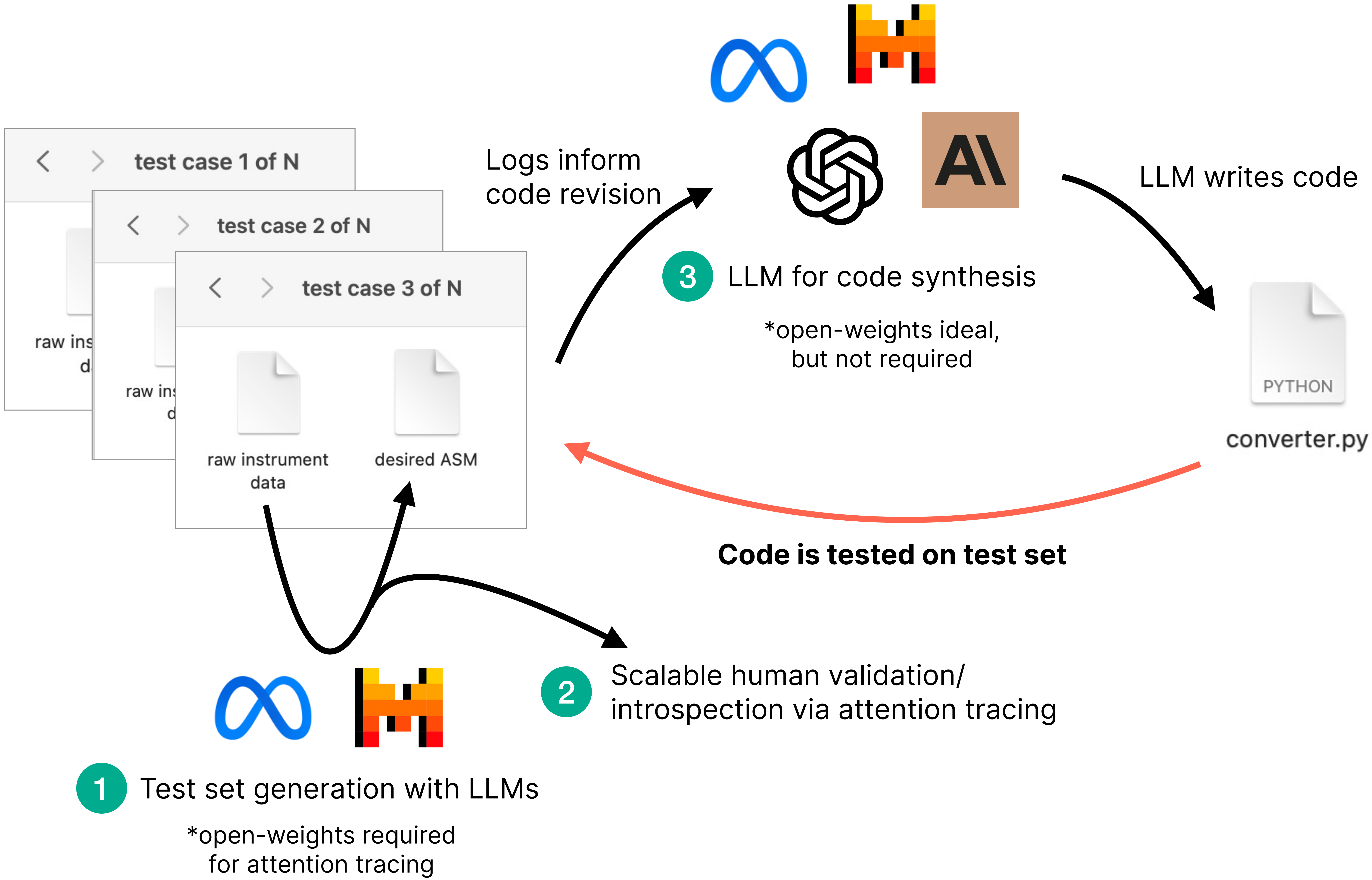
2 Scalable human validation/ introspection via attention tracing

1 Test set generation with LLMs

*open-weights required for attention tracing

Our approach

Take advantage of the good and mitigate the bad



Logs inform code revision

3 LLM for code synthesis

*open-weights ideal, but not required

LLM writes code

Code is tested on test set

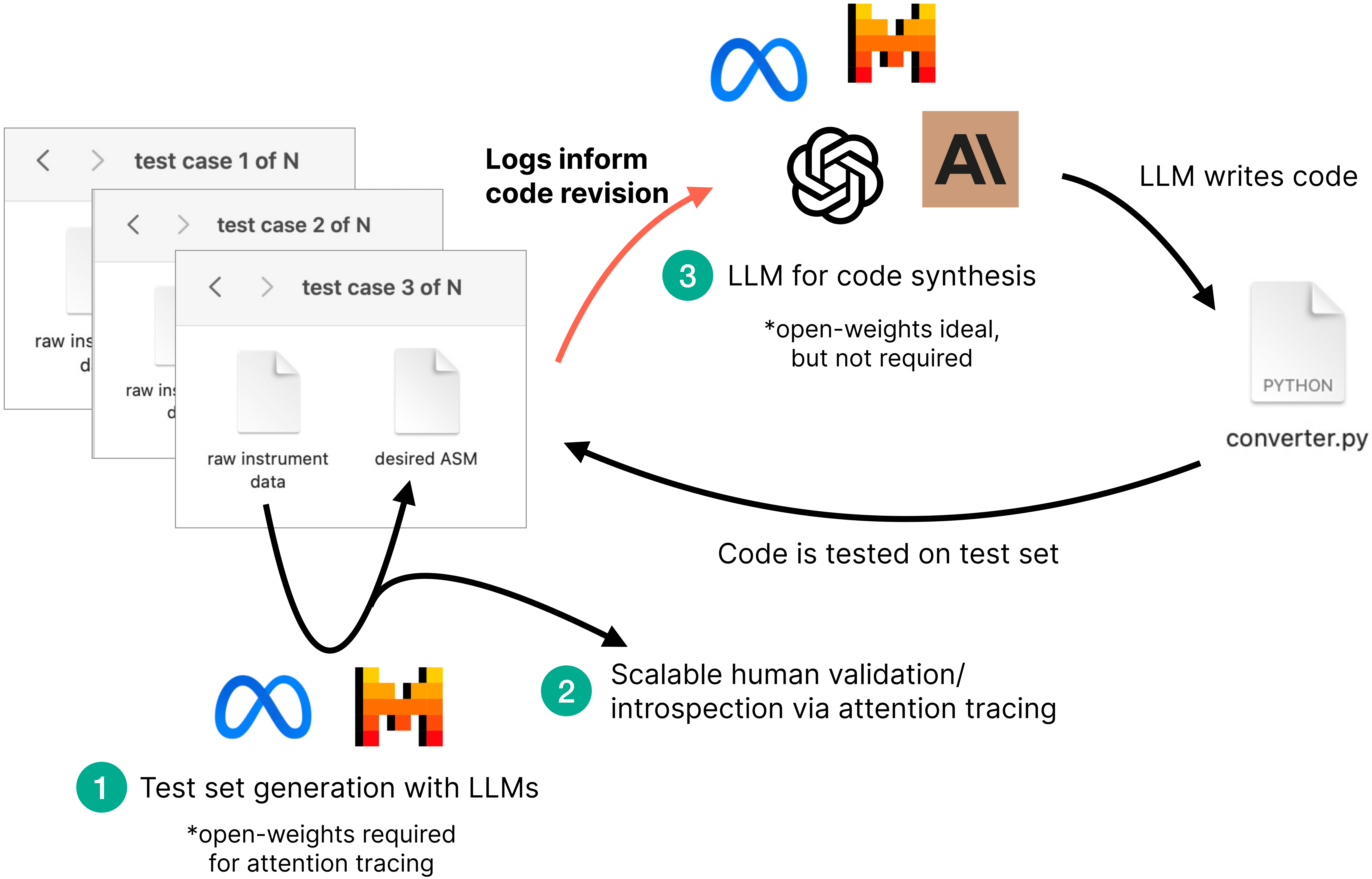
2 Scalable human validation/ introspection via attention tracing

1 Test set generation with LLMs

*open-weights required for attention tracing

Our approach

Take advantage of the good and mitigate the bad



Logs inform code revision

3 LLM for code synthesis

*open-weights ideal, but not required

LLM writes code

Code is tested on test set

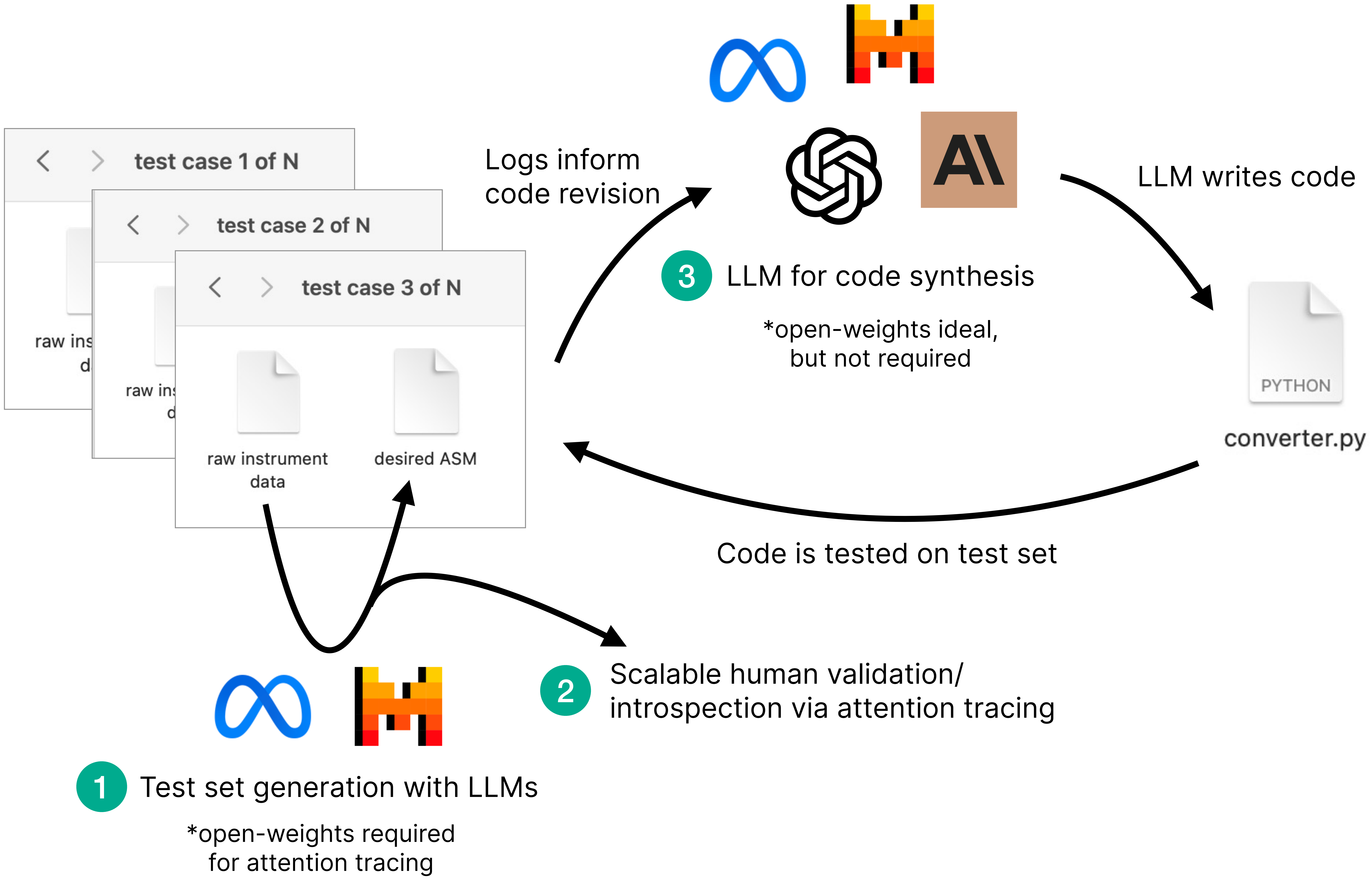
2 Scalable human validation/ introspection via attention tracing

1 Test set generation with LLMs

*open-weights required for attention tracing

Our approach

Take advantage of the good and mitigate the bad



Logs inform code revision

LLM writes code

3 LLM for code synthesis

*open-weights ideal, but not required

Code is tested on test set

2 Scalable human validation/ introspection via attention tracing

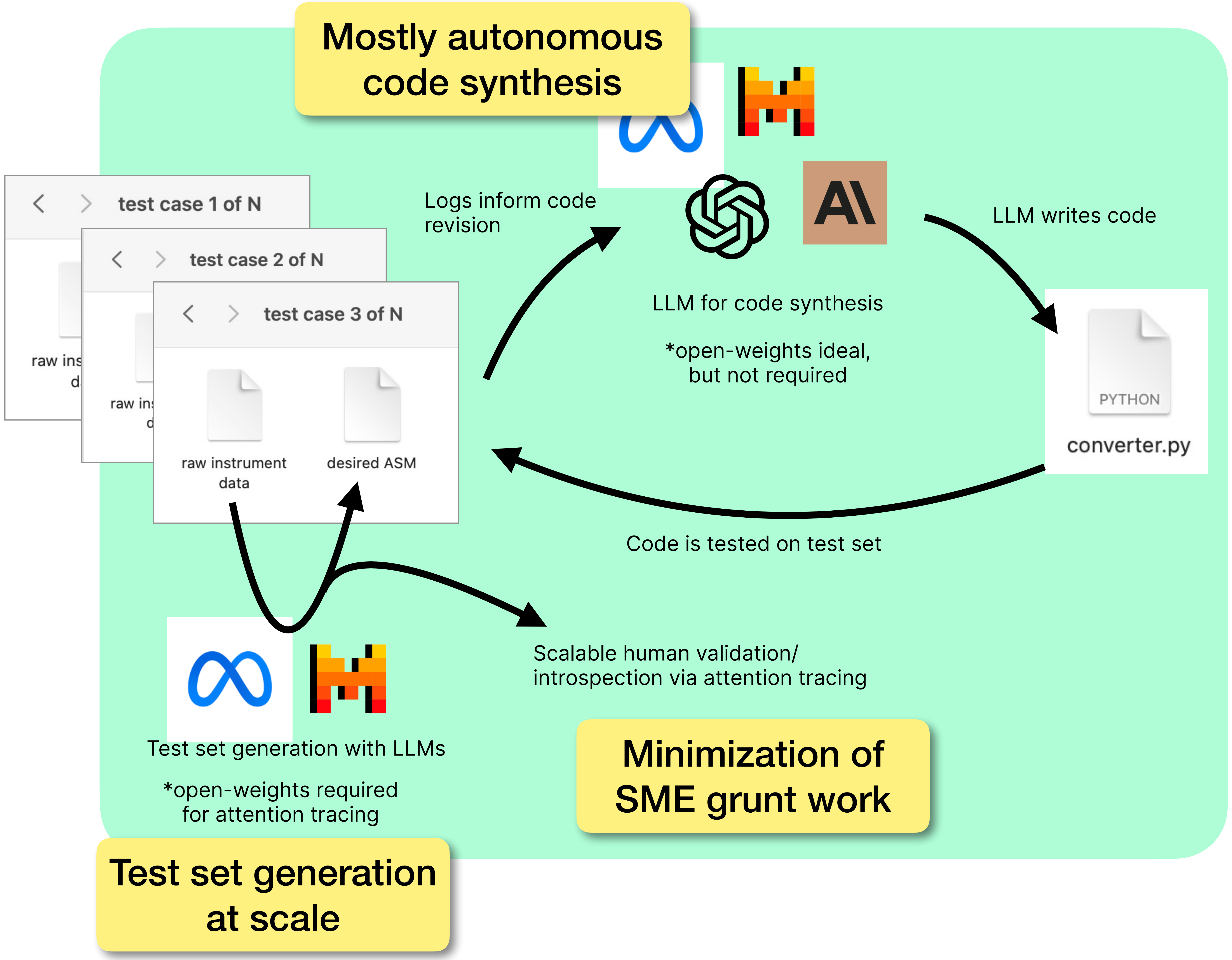
1 Test set generation with LLMs

*open-weights required for attention tracing

Why is this better and faster?

Why is this better and faster?

- **LLM-based workflow enables 1-2 day turnaround**



Mostly autonomous code synthesis

Logs inform code revision

LLM writes code

LLM for code synthesis

*open-weights ideal, but not required

Code is tested on test set

Scalable human validation/ introspection via attention tracing

Test set generation with LLMs

*open-weights required for attention tracing

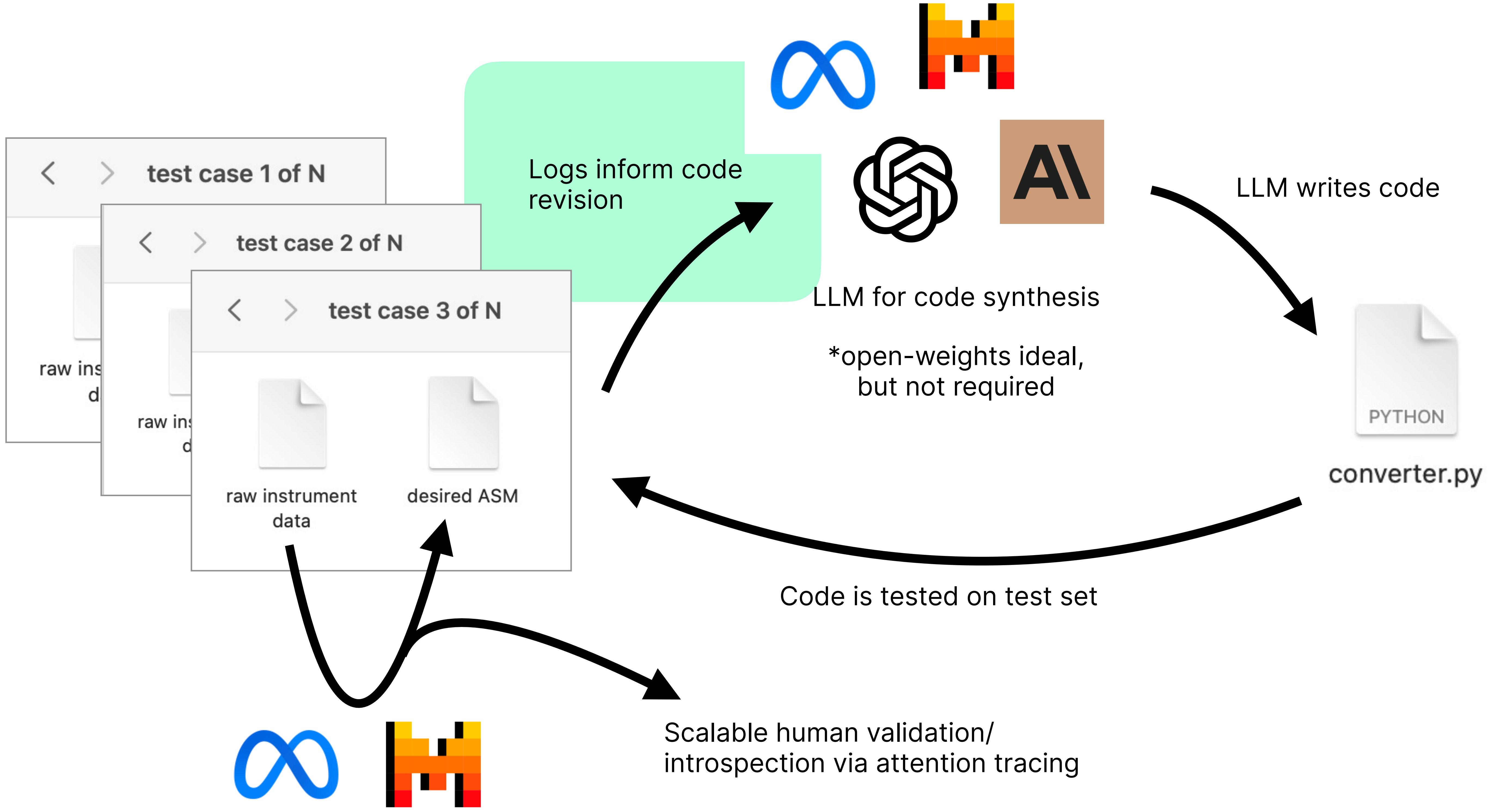
Minimization of SME grunt work

Test set generation at scale

Why is this better and faster?

- LLM-based workflow enables 1-2 day turnaround

- **LLMs can generate 100x the tests vs software engineers**



Logs inform code revision

LLM for code synthesis

*open-weights ideal, but not required

LLM writes code

converter.py

Code is tested on test set

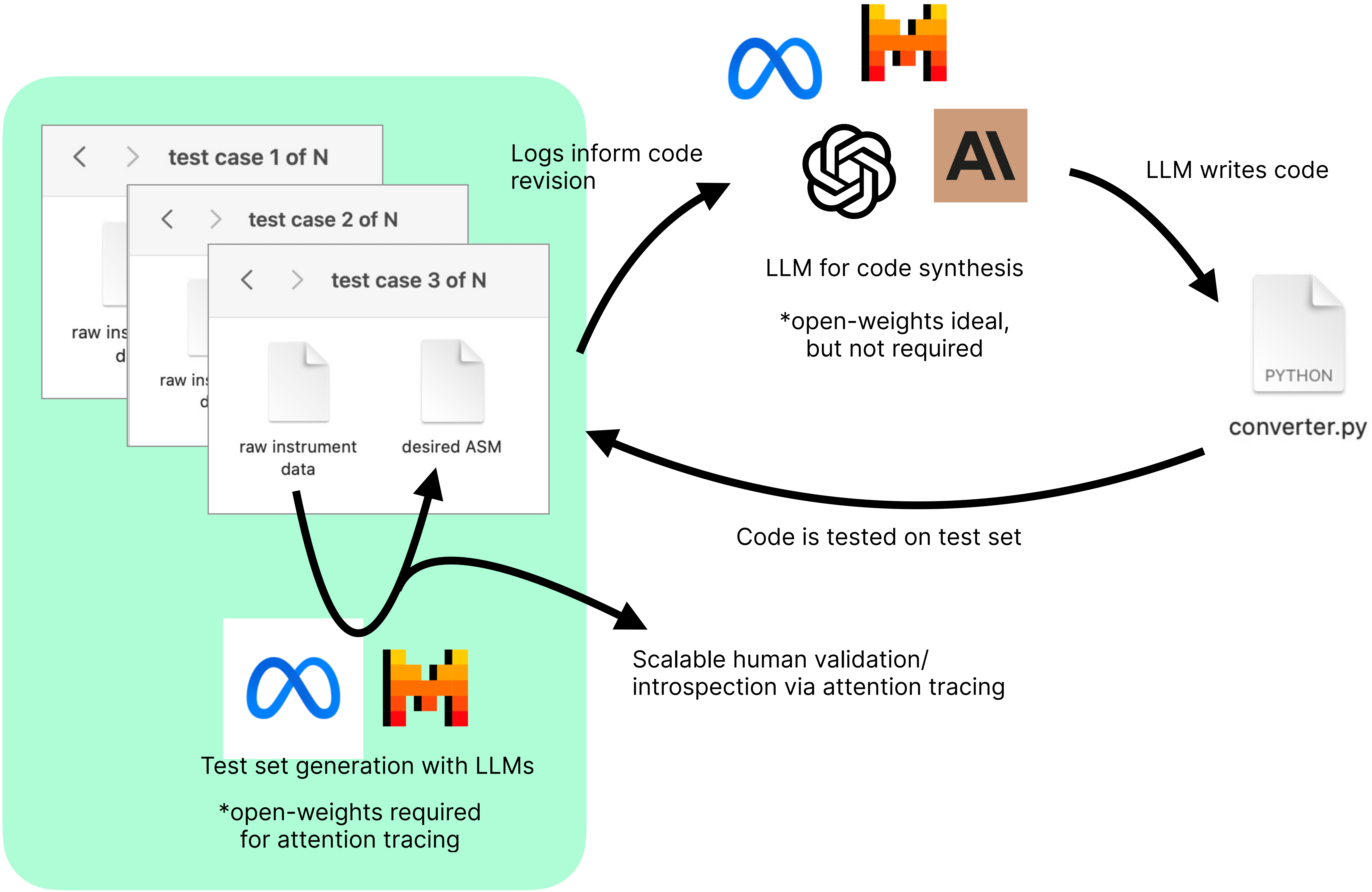
Scalable human validation/ introspection via attention tracing

Test set generation with LLMs

*open-weights required for attention tracing

Why is this better and faster?

- LLM-based workflow enables 1-2 day turnaround
- LLMs can generate 100x the tests vs software engineers
- **Take advantage of existing data for highly comprehensive testing**



Logs inform code revision

LLM for code synthesis

*open-weights ideal, but not required

LLM writes code

Code is tested on test set

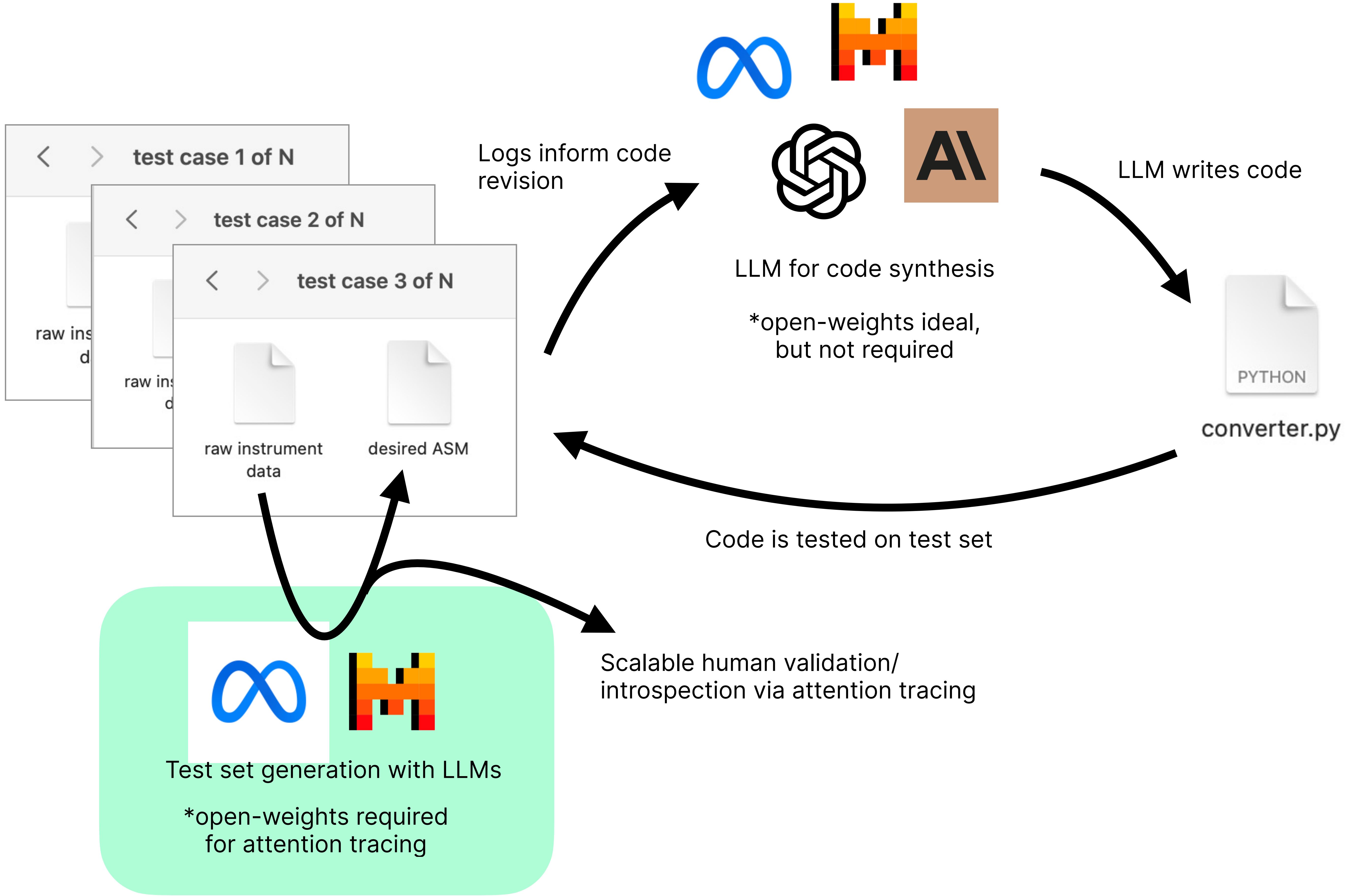
Scalable human validation/ introspection via attention tracing

Test set generation with LLMs

*open-weights required for attention tracing

Why is this better and faster?

- LLM-based workflow enables 1-2 day turnaround
- LLMs can generate 100x the tests vs software engineers
- Take advantage of existing data for highly comprehensive testing
- **LLM domain knowledge eliminates SME and engineer back-and-forth**



Logs inform code revision

LLM for code synthesis

*open-weights ideal, but not required

LLM writes code

Code is tested on test set

Scalable human validation/ introspection via attention tracing

Test set generation with LLMs

*open-weights required for attention tracing

Why is this better and faster?

- LLM-based workflow enables 1-2 day turnaround

- LLMs can generate 100x th tests vs software engineer

- Take advantage of existing data for highly comprehens testing

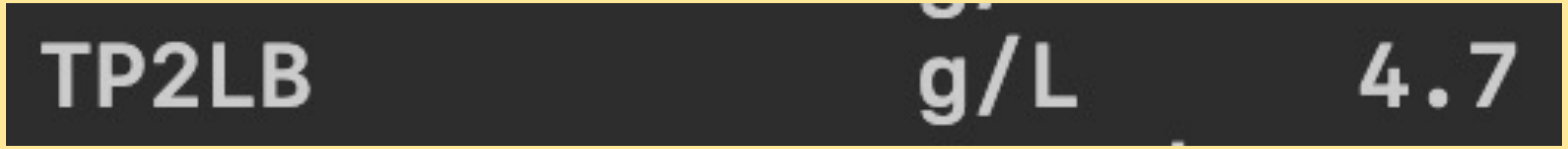
- **LLM domain knowledge eliminates SME and engin back-and-forth**

ode

THON

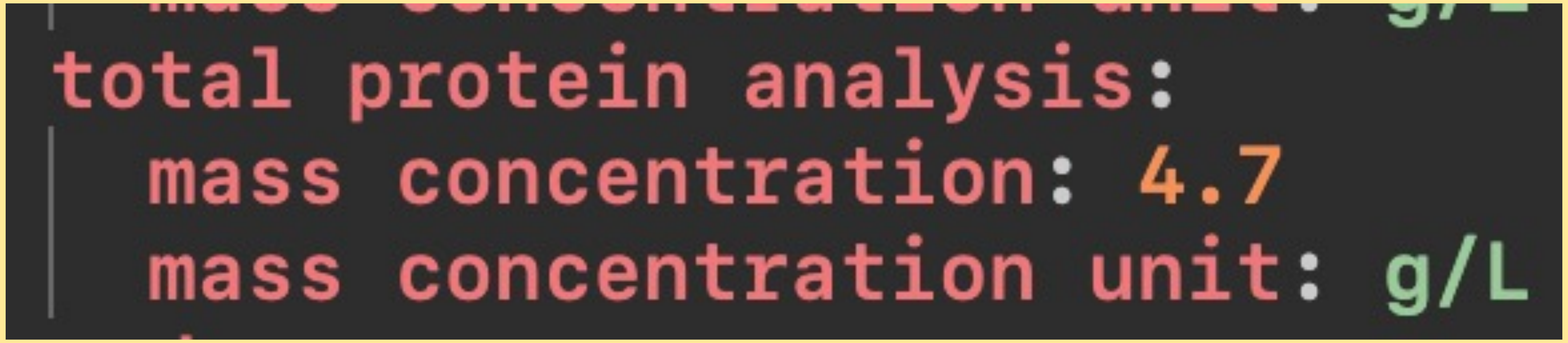
erter.py

Another benefit of the domain knowledge…



```
TP2LB                          g/L        4.7
```

Raw instrument data



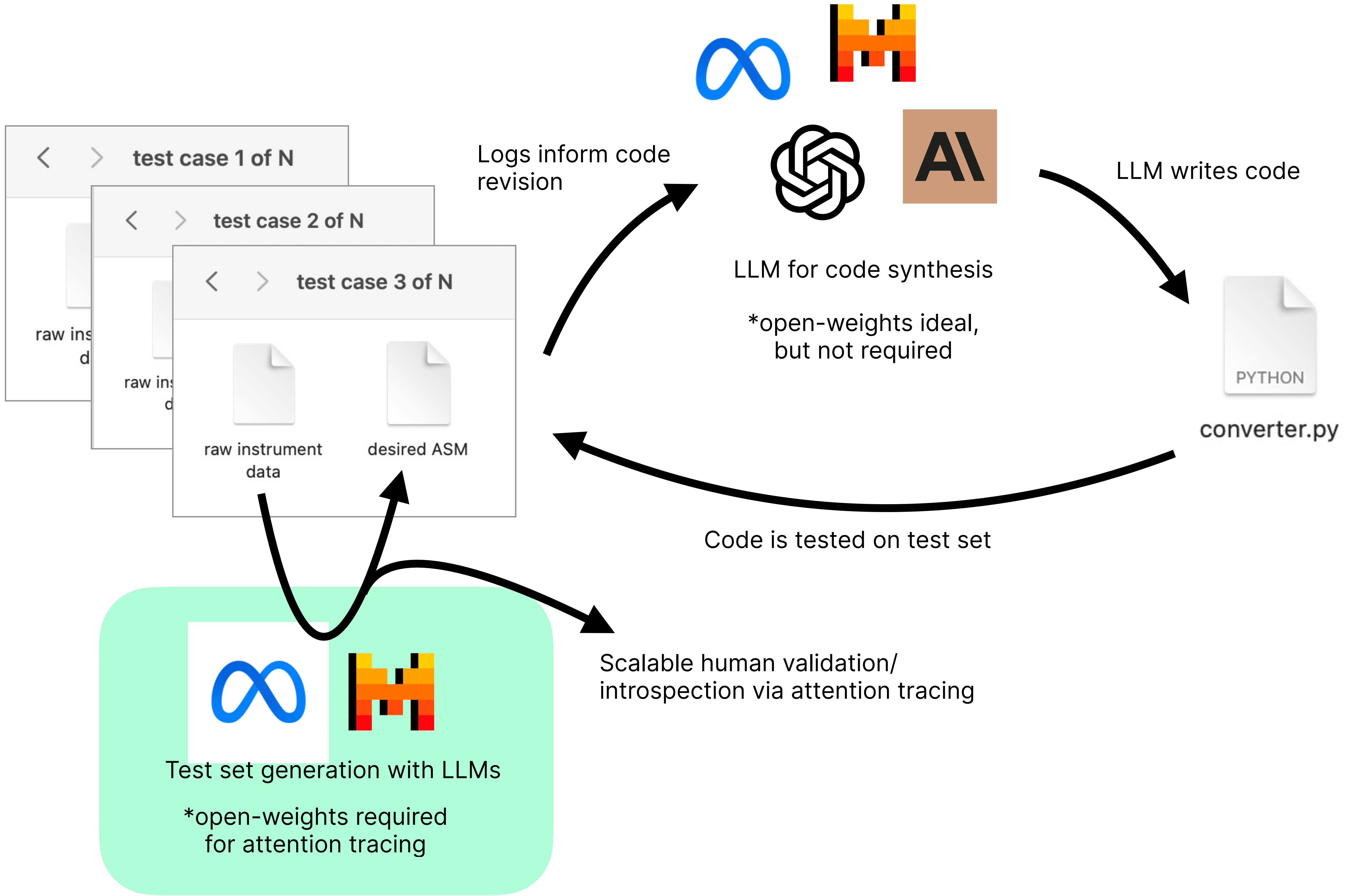
```
total protein analysis:
  mass concentration: 4.7
  mass concentration unit: g/L
```

Successful extraction of field not in ASM schema

<u>Opportunity to identify additions to Allotrope standards!</u>

Why is this better and faster?

- LLM-based workflow enables 1-2 day turnaround
- LLMs can generate 100x the tests vs software engineers
- Take advantage of existing data for highly comprehensive testing
- **LLM domain knowledge eliminates SME and engineer back-and-forth**



Logs inform code revision

LLM for code synthesis

*open-weights ideal, but not required

LLM writes code

Code is tested on test set

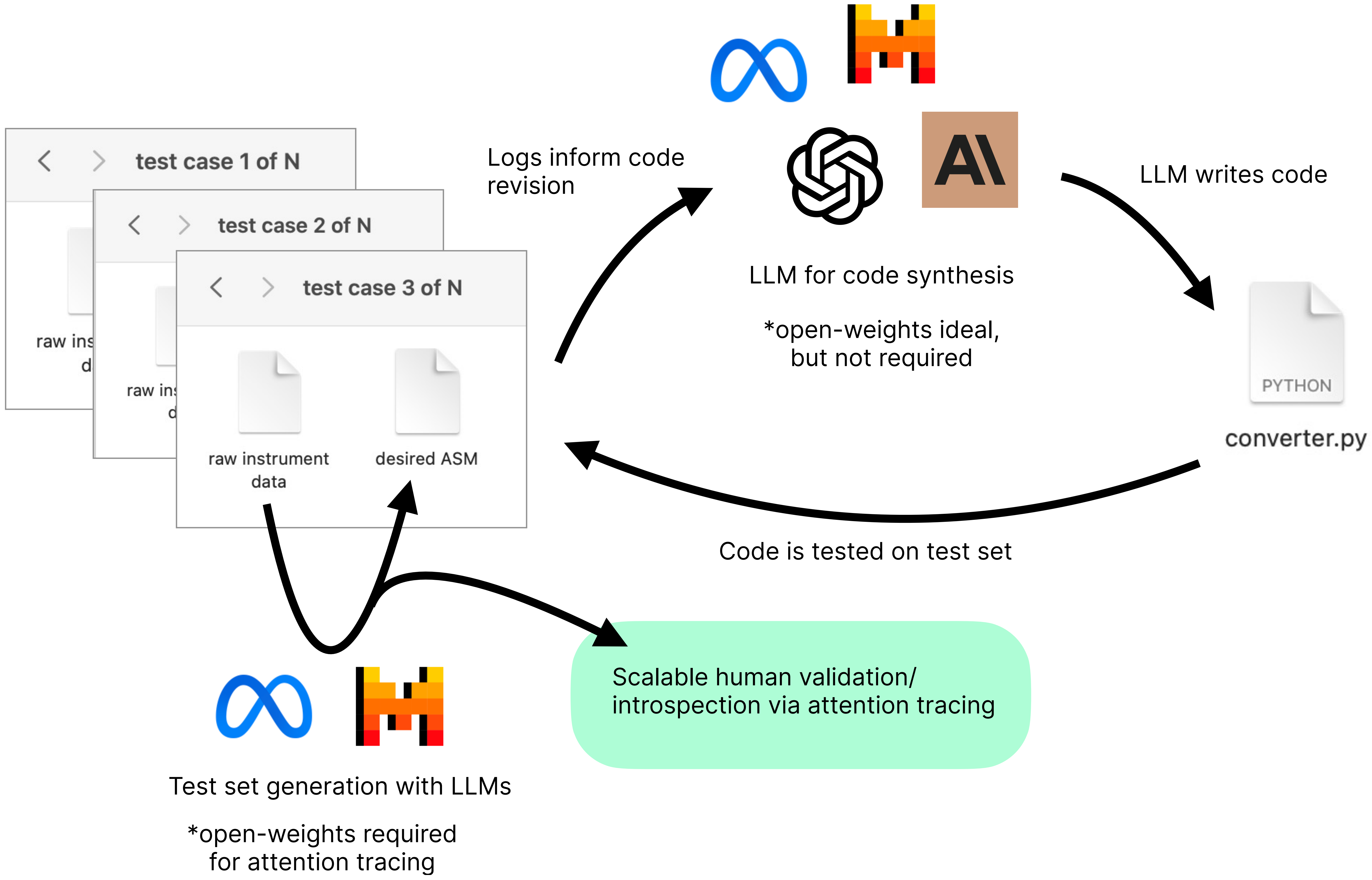
Scalable human validation/ introspection via attention tracing

Test set generation with LLMs

*open-weights required for attention tracing

Why is this better and faster?

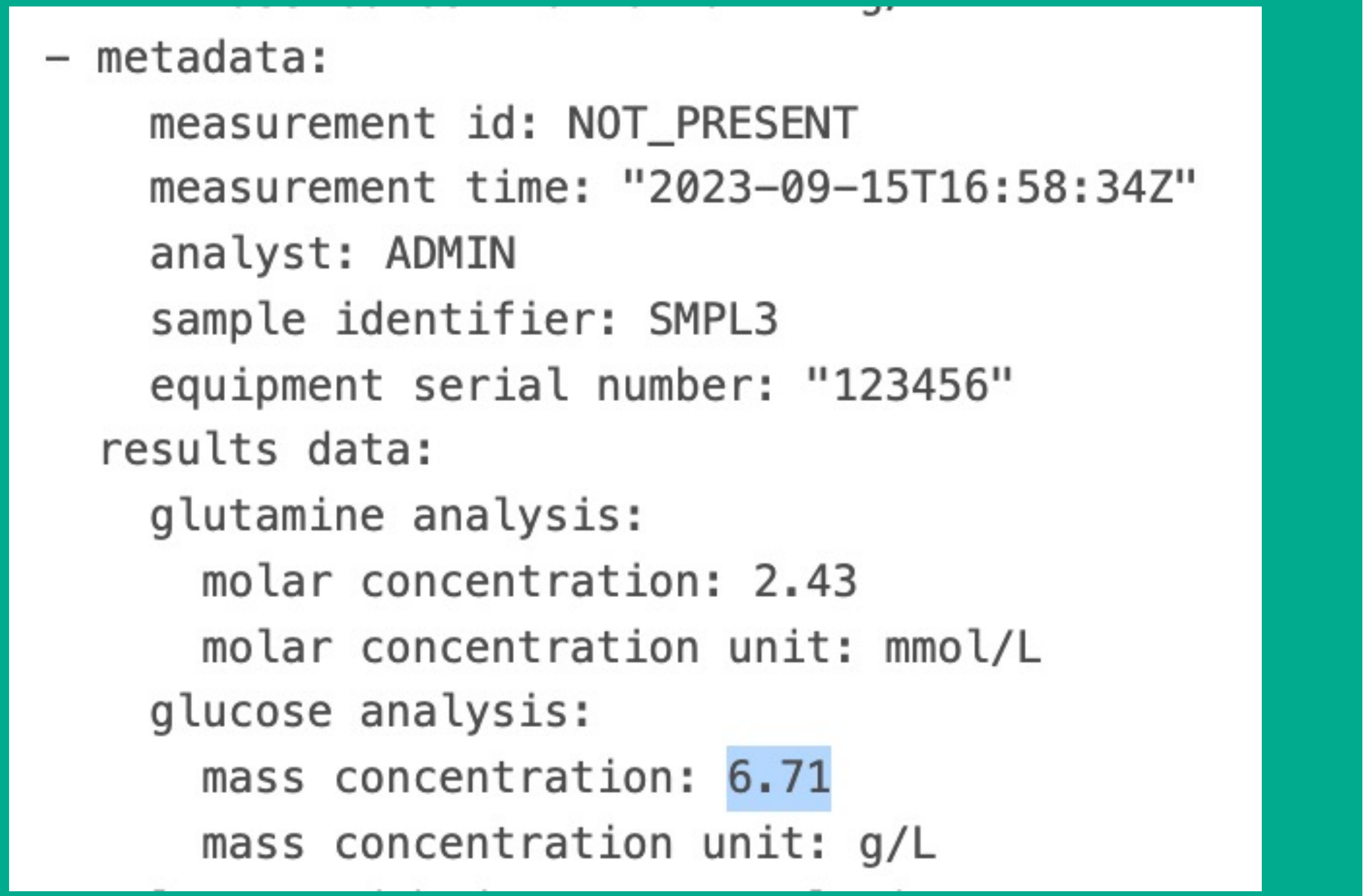
- LLM-based workflow enables 1-2 day turnaround
- LLMs can generate 100x the tests vs software engineers
- Take advantage of existing data for highly comprehensive testing
- LLM domain knowledge eliminates SME and engineer back-and-forth
- **Tooling built around LLM attention greatly speeds up manual SME validation work**



test case 1 of N
raw ins
d
test case 2 of N
raw ins
d
test case 3 of N
raw instrument data
desired ASM
Logs inform code revision
LLM for code synthesis
*open-weights ideal, but not required
LLM writes code
PYTHON
converter.py
Code is tested on test set
Scalable human validation/ introspection via attention tracing
Test set generation with LLMs
*open-weights required for attention tracing

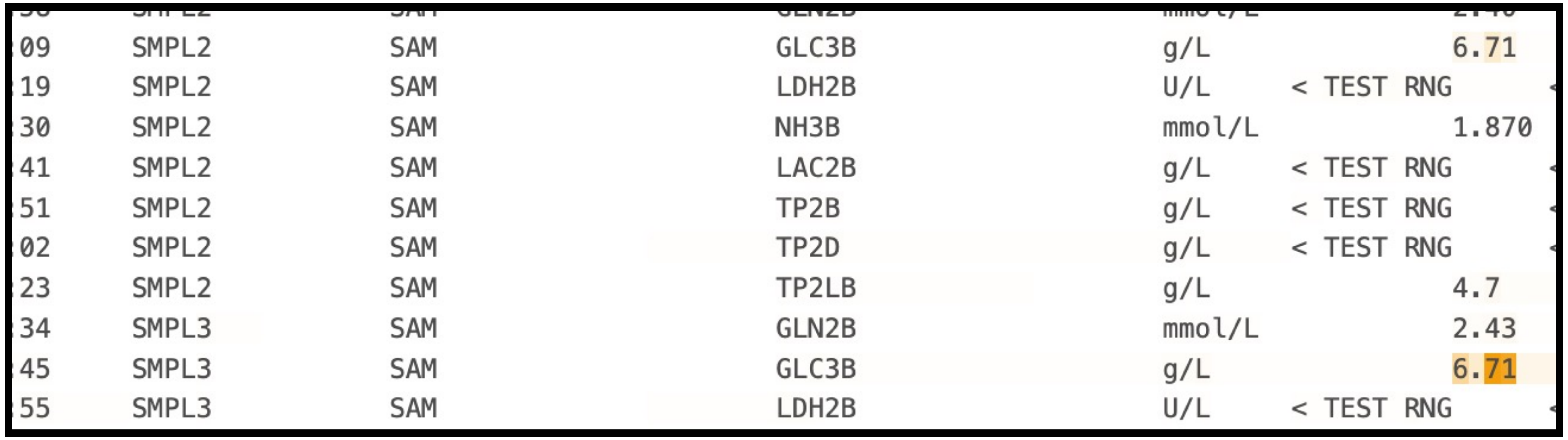
Scalable human validation/introspection via attention tracing

Snapshots from our attention tracing tool



```
- metadata:
    measurement id: NOT_PRESENT
    measurement time: "2023-09-15T16:58:34Z"
    analyst: ADMIN
    sample identifier: SMPL3
    equipment serial number: "123456"
  results data:
    glutamine analysis:
      molar concentration: 2.43
      molar concentration unit: mmol/L
    glucose analysis:
      mass concentration: 6.71
      mass concentration unit: g/L
```

Highlighting LLM interpretation in our attention tracing tool



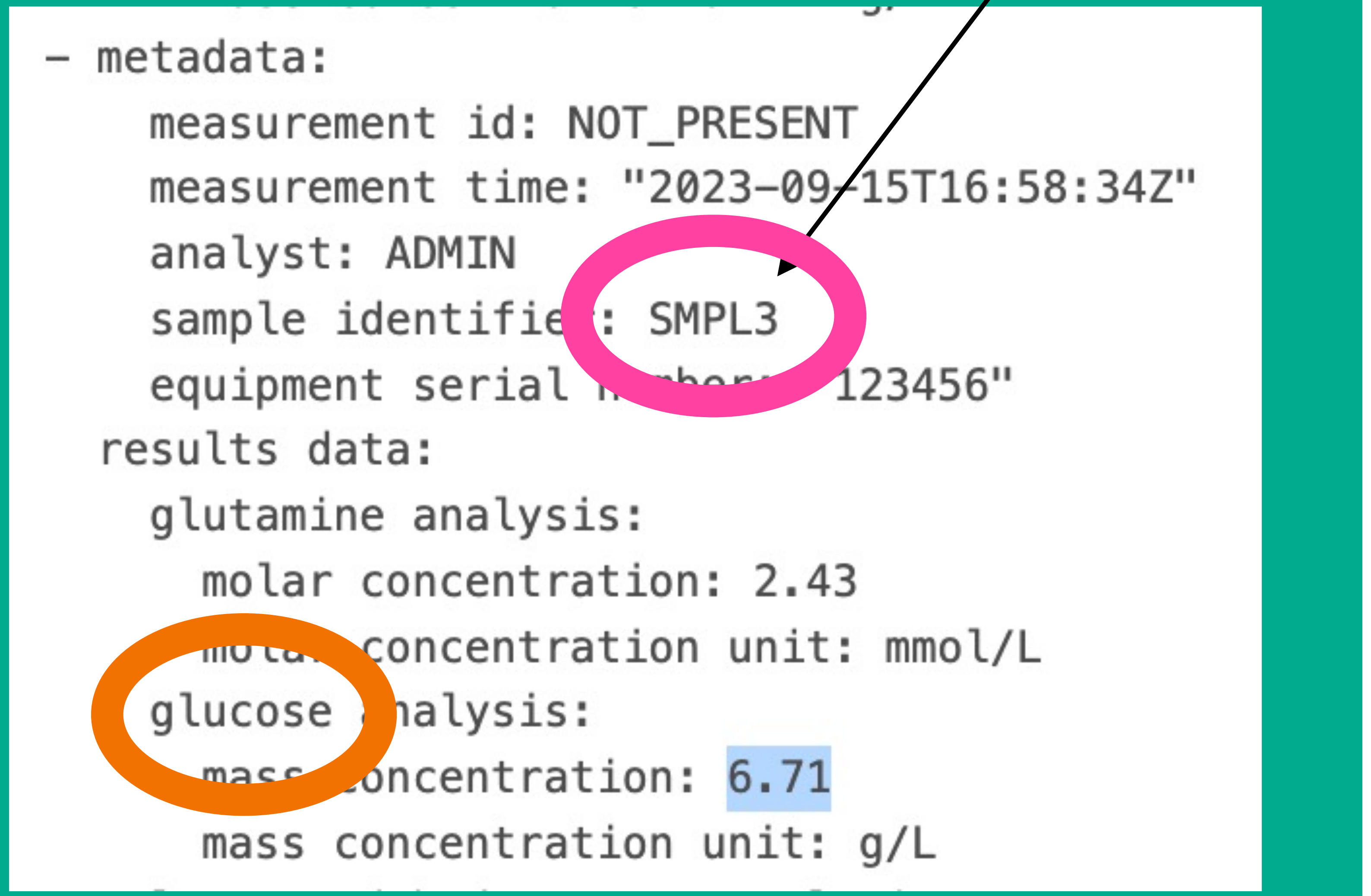
```
09    SMPL2      SAM          GLC3B         g/L               6.71
19    SMPL2      SAM          LDH2B         U/L       < TEST RNG
30    SMPL2      SAM          NH3B          mmol/L            1.870
41    SMPL2      SAM          LAC2B         g/L       < TEST RNG
51    SMPL2      SAM          TP2B          g/L       < TEST RNG
02    SMPL2      SAM          TP2D          g/L       < TEST RNG
23    SMPL2      SAM          TP2LB         g/L               4.7
34    SMPL3      SAM          GLN2B         mmol/L            2.43
45    SMPL3      SAM          GLC3B         g/L               6.71
55    SMPL3      SAM          LDH2B         U/L       < TEST RNG
```

Raw instrument data

Scalable human validation/introspection via attention tracing

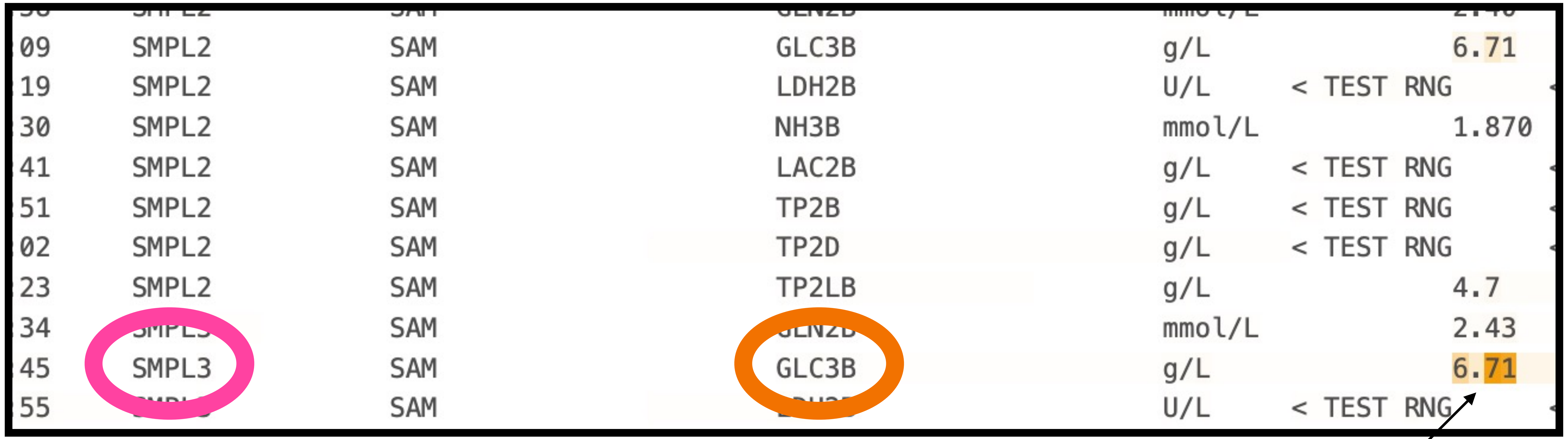
Snapshots from our attention tracing tool

This glucose result is from SMPL3



```
- metadata:
    measurement id: NOT_PRESENT
    measurement time: "2023-09-15T16:58:34Z"
    analyst: ADMIN
    sample identifier: SMPL3
    equipment serial number: "123456"
  results data:
    glutamine analysis:
      molar concentration: 2.43
      molar concentration unit: mmol/L
    glucose analysis:
      mass concentration: 6.71
      mass concentration unit: g/L
```

Highlighting LLM interpretation in our attention tracing tool



09	SMPL2	SAM	GLC3B	g/L		6.71
19	SMPL2	SAM	LDH2B	U/L	< TEST RNG	
30	SMPL2	SAM	NH3B	mmol/L		1.870
41	SMPL2	SAM	LAC2B	g/L	< TEST RNG	
51	SMPL2	SAM	TP2B	g/L	< TEST RNG	
02	SMPL2	SAM	TP2D	g/L	< TEST RNG	
23	SMPL2	SAM	TP2LB	g/L		4.7
34	SMPL3	SAM	GLN2B	mmol/L		2.43
45	SMPL3	SAM	GLC3B	g/L		6.71
55	SMPL3	SAM	LDH2B	U/L	< TEST RNG	

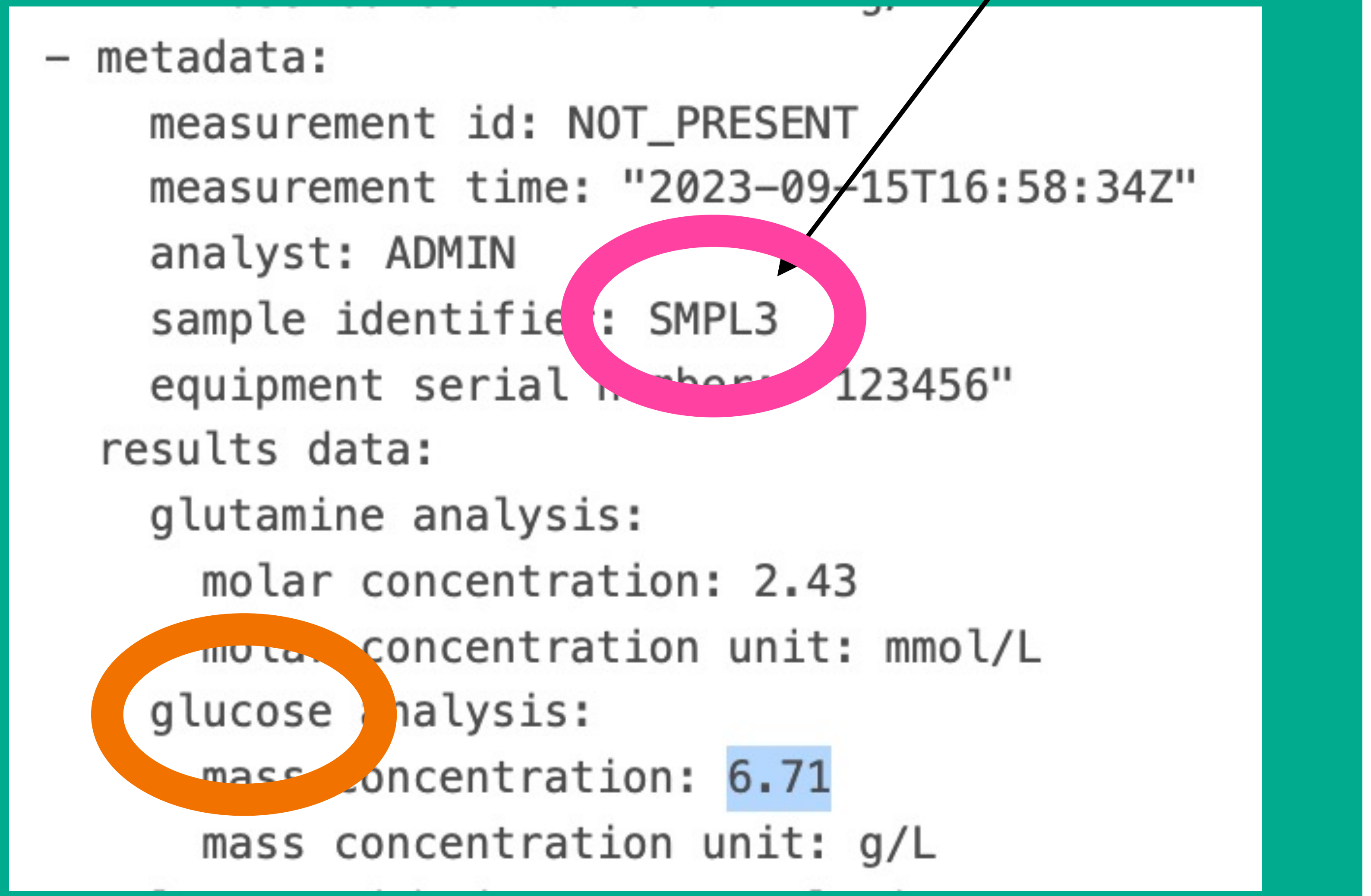
Raw instrument data

Attention is only directed at the correct 6.71 from SMPL3

Scalable human validation/introspection via attention tracing

Snapshots from our attention tracing tool

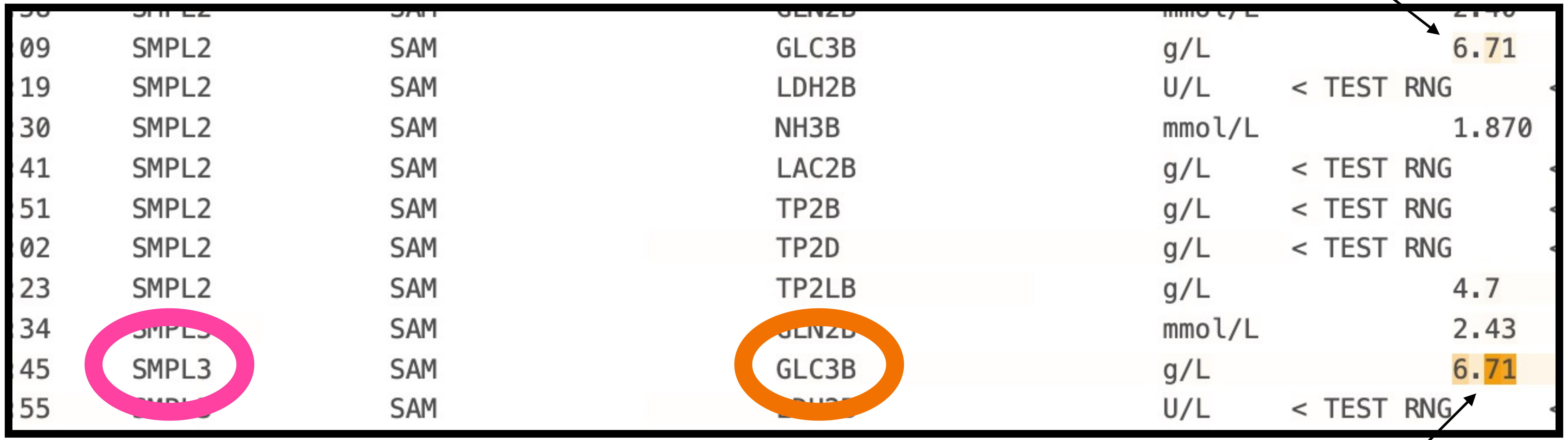
This glucose result is from SMPL3



```
- metadata:
    measurement id: NOT_PRESENT
    measurement time: "2023-09-15T16:58:34Z"
    analyst: ADMIN
    sample identifie : SMPL3
    equipment serial n        123456"
  results data:
    glutamine analysis:
      molar concentration: 2.43
      mo a  concentration unit: mmol/L
    glucose  nalysis:
      mas  oncentration: 6.71
      mass concentration unit: g/L
```

Highlighting LLM interpretation in our attention tracing tool

Also Glucose, also 6.71, but from a different sample (SMPL2)



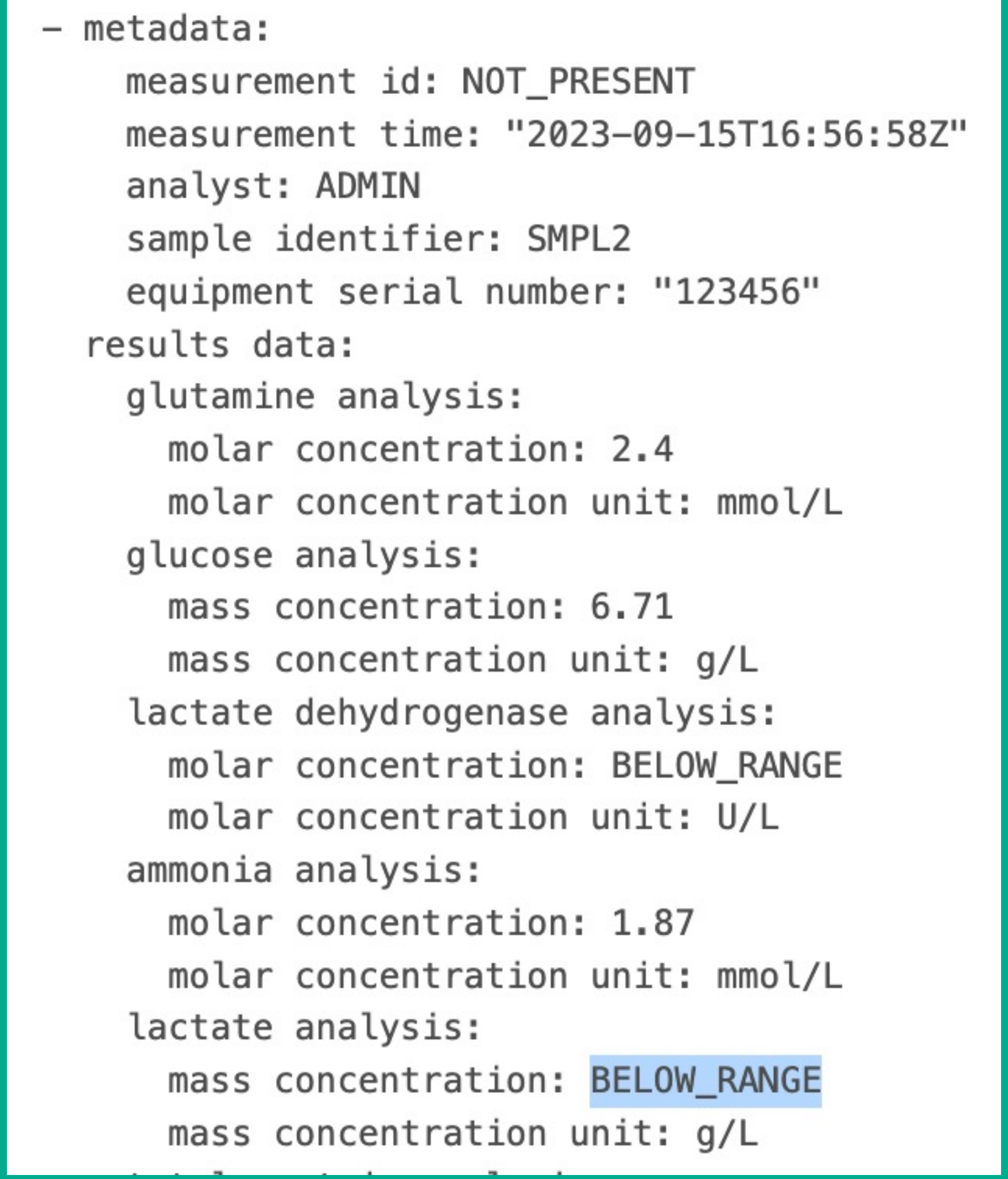
```
09     SMPL2         SAM                GLC3B         g/L              6.71
19     SMPL2         SAM                LDH2B         U/L     < TEST RNG
30     SMPL2         SAM                NH3B          mmol/L           1.870
41     SMPL2         SAM                LAC2B         g/L     < TEST RNG
51     SMPL2         SAM                TP2B          g/L     < TEST RNG
02     SMPL2         SAM                TP2D          g/L     < TEST RNG
23     SMPL2         SAM                TP2LB         g/L              4.7
34     SMPL3         SAM                GLN2B         mmol/L           2.43
45     SMPL3         SAM                GLC3B         g/L              6.71
55                   SAM                              U/L     < TEST RNG
```

Raw instrument data

Attention is only directed at the correct 6.71 from SMPL3

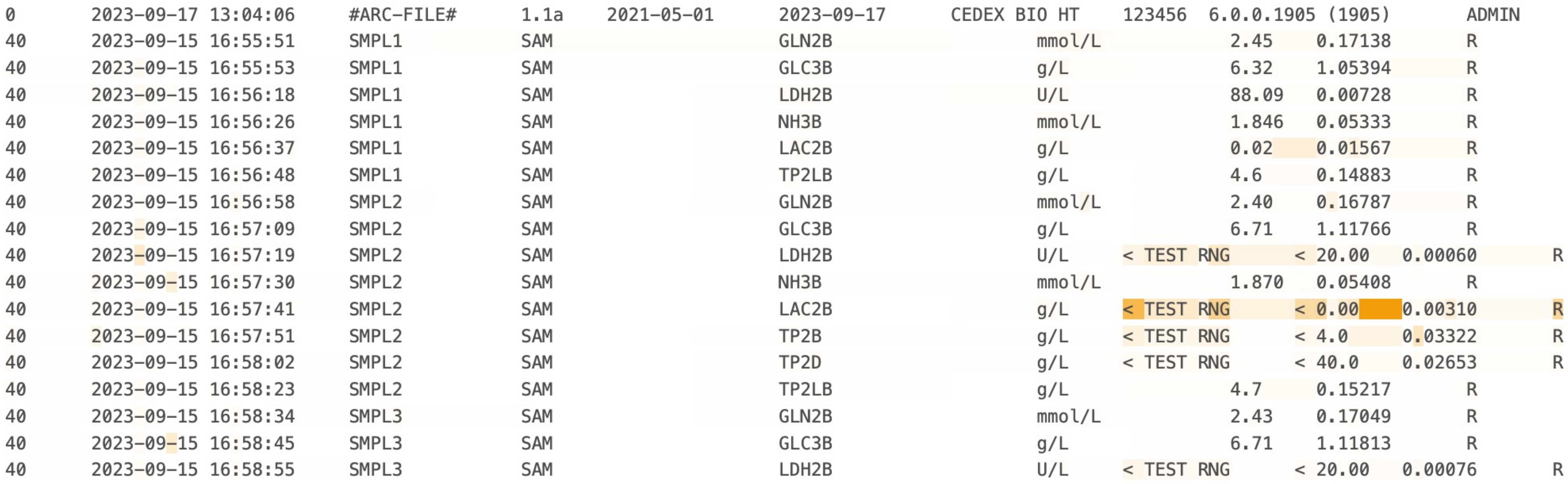
Scalable human validation/introspection via attention tracing

Snapshots from our attention tracing tool



```
- metadata:
    measurement id: NOT_PRESENT
    measurement time: "2023-09-15T16:56:58Z"
    analyst: ADMIN
    sample identifier: SMPL2
    equipment serial number: "123456"
  results data:
    glutamine analysis:
      molar concentration: 2.4
      molar concentration unit: mmol/L
    glucose analysis:
      mass concentration: 6.71
      mass concentration unit: g/L
    lactate dehydrogenase analysis:
      molar concentration: BELOW_RANGE
      molar concentration unit: U/L
    ammonia analysis:
      molar concentration: 1.87
      molar concentration unit: mmol/L
    lactate analysis:
      mass concentration: BELOW_RANGE
      mass concentration unit: g/L
```

Highlighting LLM interpretation in our attention tracing tool

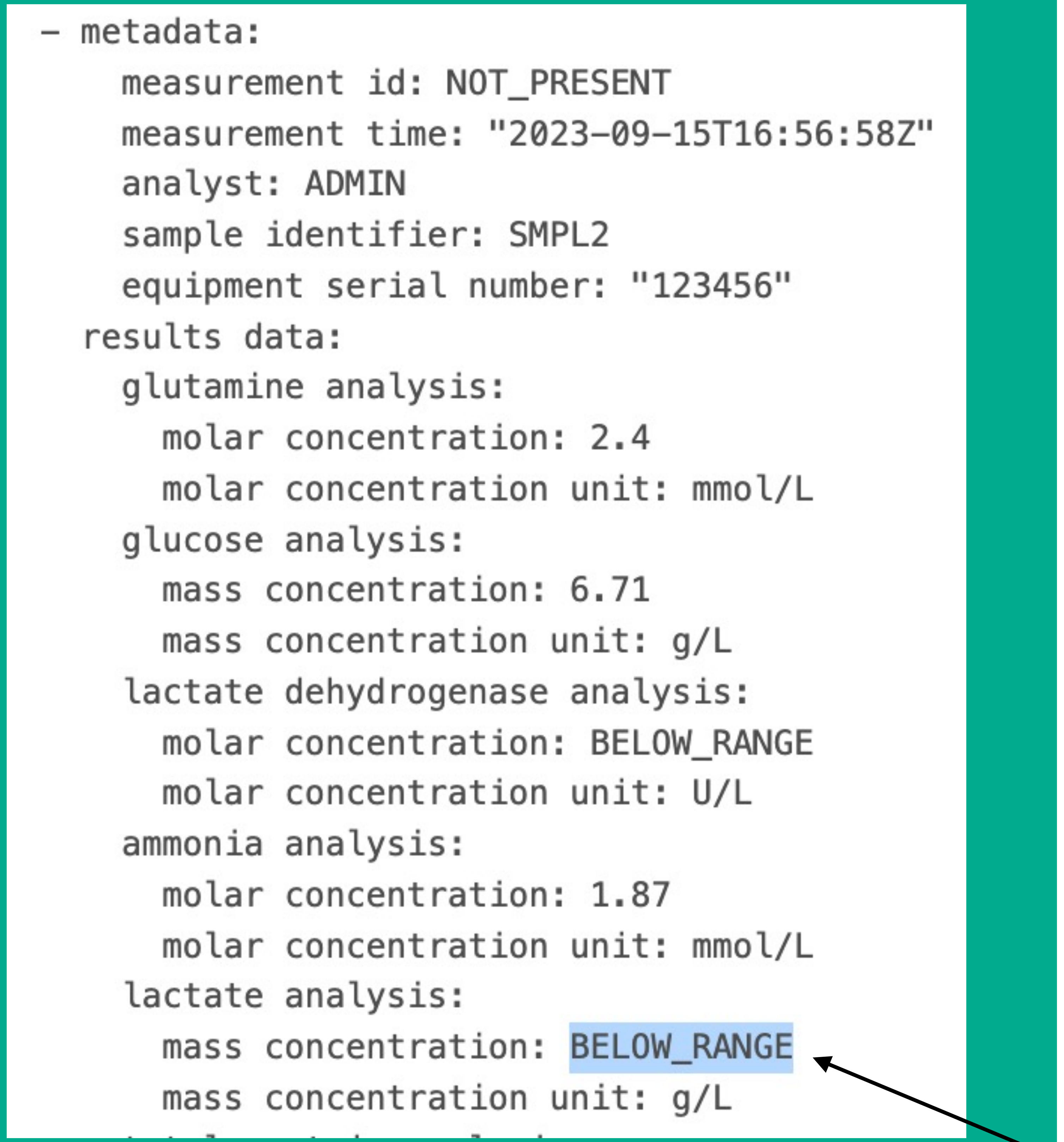


```
0     2023-09-17 13:04:06   #ARC-FILE#   1.1a   2021-05-01   2023-09-17   CEDEX BIO HT   123456   6.0.0.1905 (1905)      ADMIN
40    2023-09-15 16:55:51   SMPL1        SAM                 GLN2B                 mmol/L                  2.45    0.17138      R
40    2023-09-15 16:55:53   SMPL1        SAM                 GLC3B                 g/L                     6.32    1.05394      R
40    2023-09-15 16:56:18   SMPL1        SAM                 LDH2B                 U/L                     88.09   0.00728      R
40    2023-09-15 16:56:26   SMPL1        SAM                 NH3B                  mmol/L                  1.846   0.05333      R
40    2023-09-15 16:56:37   SMPL1        SAM                 LAC2B                 g/L                     0.02    0.01567      R
40    2023-09-15 16:56:48   SMPL1        SAM                 TP2LB                 g/L                     4.6     0.14883      R
40    2023-09-15 16:56:58   SMPL2        SAM                 GLN2B                 mmol/L                  2.40    0.16787      R
40    2023-09-15 16:57:09   SMPL2        SAM                 GLC3B                 g/L                     6.71    1.11766      R
40    2023-09-15 16:57:19   SMPL2        SAM                 LDH2B                 U/L       < TEST RNG      < 20.00    0.00060       R
40    2023-09-15 16:57:30   SMPL2        SAM                 NH3B                  mmol/L                  1.870   0.05408      R
40    2023-09-15 16:57:41   SMPL2        SAM                 LAC2B                 g/L       < TEST RNG      < 0.00     0.00310       R
40    2023-09-15 16:57:51   SMPL2        SAM                 TP2B                  g/L       < TEST RNG      < 4.0      0.03322       R
40    2023-09-15 16:58:02   SMPL2        SAM                 TP2D                  g/L       < TEST RNG      < 40.0     0.02653       R
40    2023-09-15 16:58:23   SMPL2        SAM                 TP2LB                 g/L                     4.7     0.15217      R
40    2023-09-15 16:58:34   SMPL3        SAM                 GLN2B                 mmol/L                  2.43    0.17049      R
40    2023-09-15 16:58:45   SMPL3        SAM                 GLC3B                 g/L                     6.71    1.11813      R
40    2023-09-15 16:58:55   SMPL3        SAM                 LDH2B                 U/L       < TEST RNG      < 20.00    0.00076       R
```

Raw instrument data

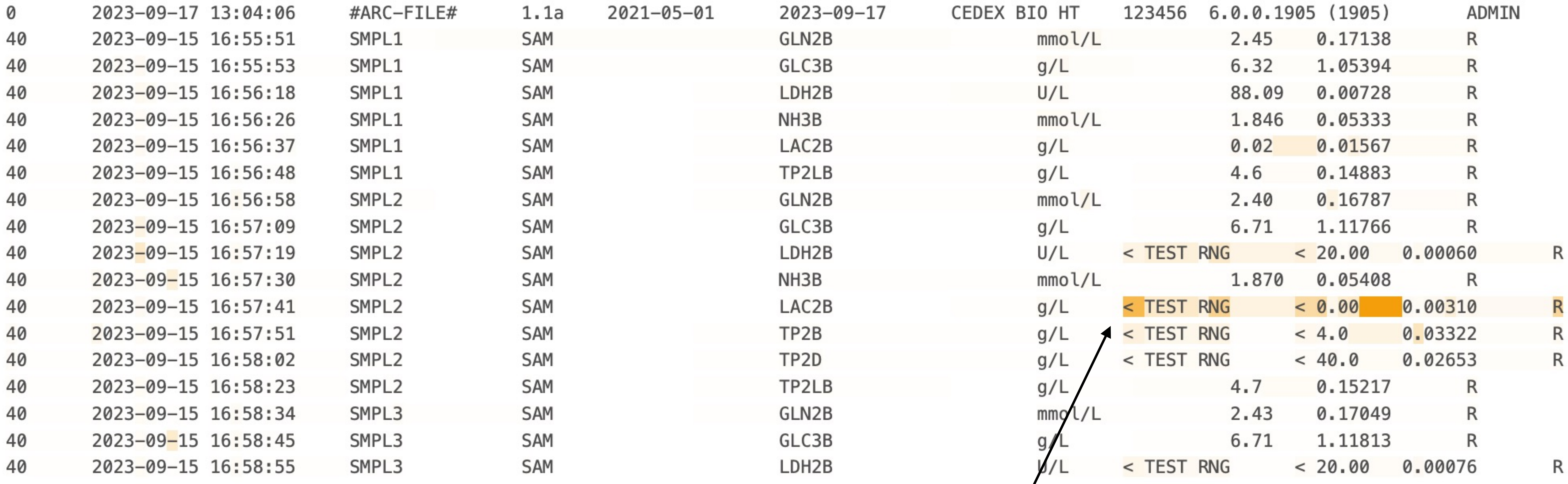
Scalable human validation/introspection via attention tracing

Snapshots from our attention tracing tool



```
- metadata:
    measurement id: NOT_PRESENT
    measurement time: "2023-09-15T16:56:58Z"
    analyst: ADMIN
    sample identifier: SMPL2
    equipment serial number: "123456"
  results data:
    glutamine analysis:
      molar concentration: 2.4
      molar concentration unit: mmol/L
    glucose analysis:
      mass concentration: 6.71
      mass concentration unit: g/L
    lactate dehydrogenase analysis:
      molar concentration: BELOW_RANGE
      molar concentration unit: U/L
    ammonia analysis:
      molar concentration: 1.87
      molar concentration unit: mmol/L
    lactate analysis:
      mass concentration: BELOW_RANGE
      mass concentration unit: g/L
```

Highlighting LLM interpretation in our attention tracing tool



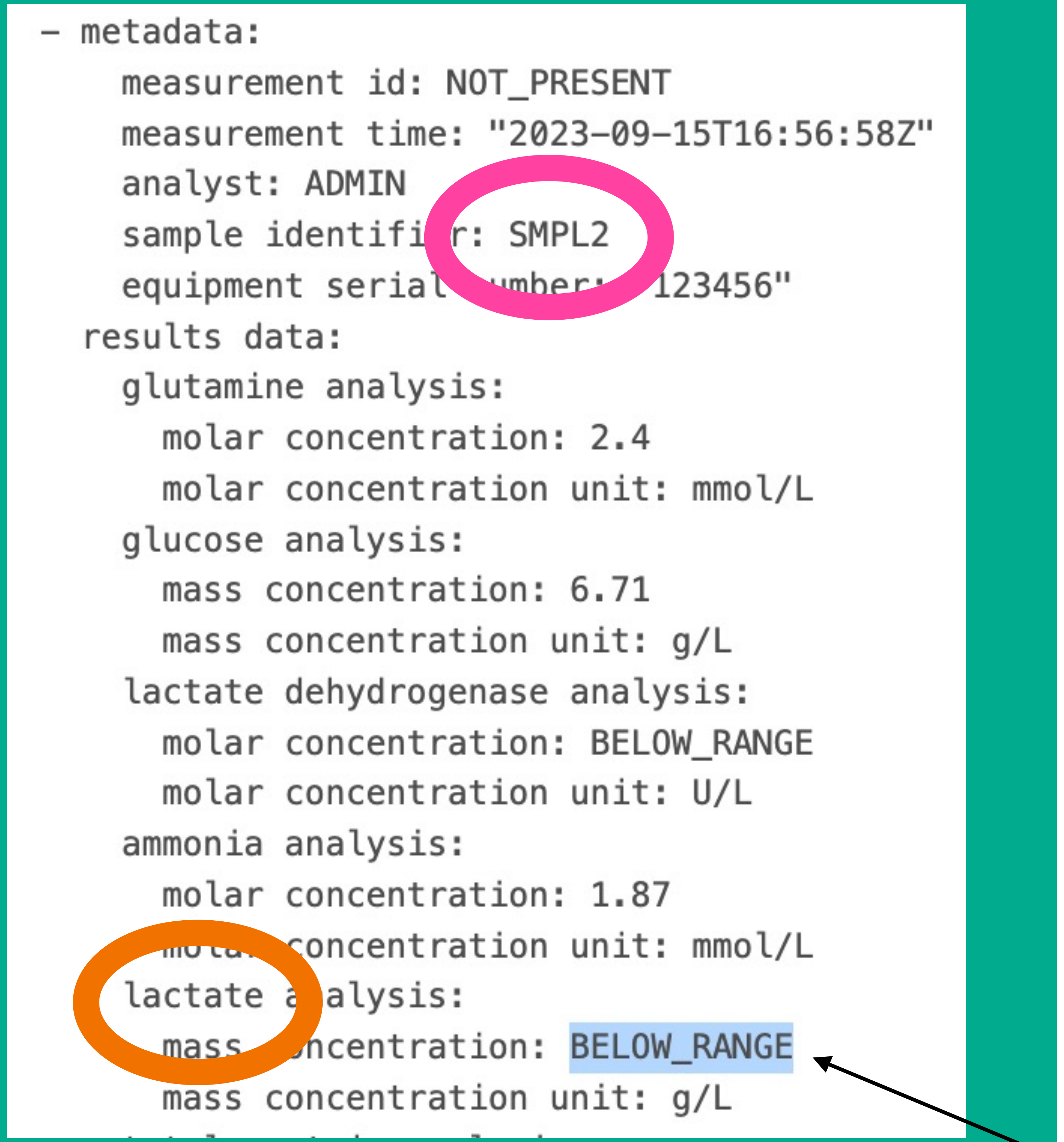
```
0     2023-09-17 13:04:06    #ARC-FILE#    1.1a    2021-05-01    2023-09-17    CEDEX BIO HT    123456  6.0.0.1905 (1905)      ADMIN
40    2023-09-15 16:55:51    SMPL1         SAM                   GLN2B               mmol/L                2.45    0.17138      R
40    2023-09-15 16:55:53    SMPL1         SAM                   GLC3B               g/L                   6.32    1.05394      R
40    2023-09-15 16:56:18    SMPL1         SAM                   LDH2B               U/L                   88.09   0.00728      R
40    2023-09-15 16:56:26    SMPL1         SAM                   NH3B                mmol/L                1.846   0.05333      R
40    2023-09-15 16:56:37    SMPL1         SAM                   LAC2B               g/L                   0.02    0.01567      R
40    2023-09-15 16:56:48    SMPL1         SAM                   TP2LB               g/L                   4.6     0.14883      R
40    2023-09-15 16:56:58    SMPL2         SAM                   GLN2B               mmol/L                2.40    0.16787      R
40    2023-09-15 16:57:09    SMPL2         SAM                   GLC3B               g/L                   6.71    1.11766      R
40    2023-09-15 16:57:19    SMPL2         SAM                   LDH2B               U/L     < TEST RNG    < 20.00    0.00060      R
40    2023-09-15 16:57:30    SMPL2         SAM                   NH3B                mmol/L                1.870   0.05408      R
40    2023-09-15 16:57:41    SMPL2         SAM                   LAC2B               g/L     < TEST RNG    < 0.00     0.00310      R
40    2023-09-15 16:57:51    SMPL2         SAM                   TP2B                g/L     < TEST RNG    < 4.0      0.03322      R
40    2023-09-15 16:58:02    SMPL2         SAM                   TP2D                g/L     < TEST RNG    < 40.0     0.02653      R
40    2023-09-15 16:58:23    SMPL2         SAM                   TP2LB               g/L                   4.7     0.15217      R
40    2023-09-15 16:58:34    SMPL3         SAM                   GLN2B               mmol/L                2.43    0.17049      R
40    2023-09-15 16:58:45    SMPL3         SAM                   GLC3B               g/L                   6.71    1.11813      R
40    2023-09-15 16:58:55    SMPL3         SAM                   LDH2B               U/L     < TEST RNG    < 20.00    0.00076      R
```

Raw instrument data

Correctly interprets `< TEST RNG` value as below measurement range

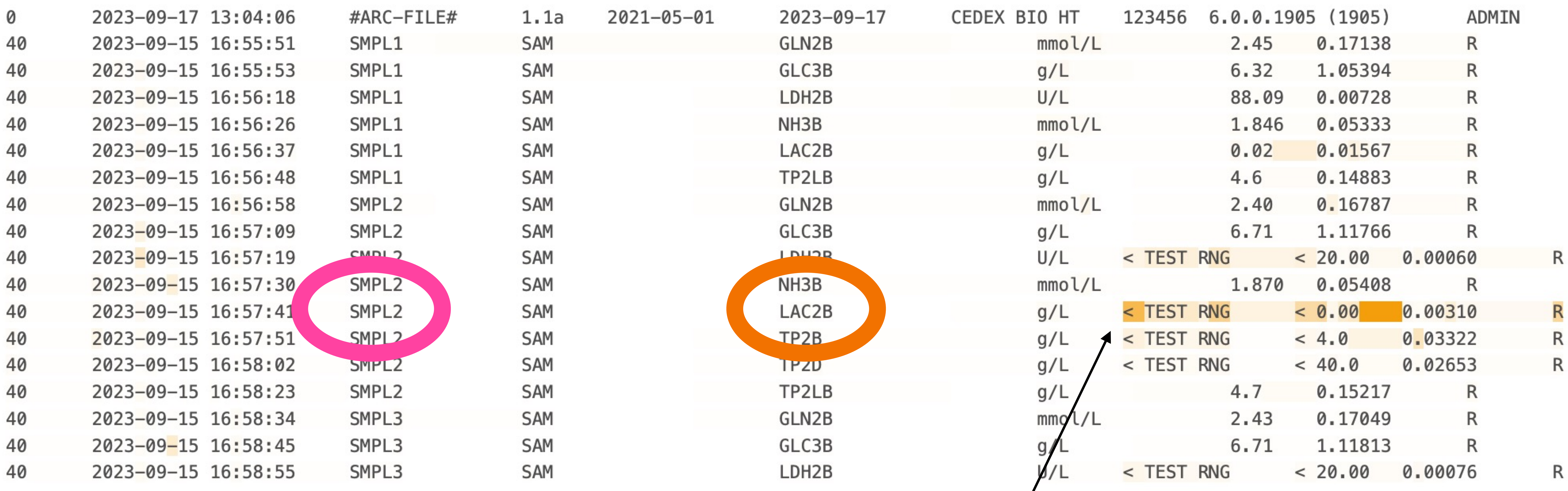
Scalable human validation/introspection via attention tracing

Snapshots from our attention tracing tool



```
- metadata:
    measurement id: NOT_PRESENT
    measurement time: "2023-09-15T16:56:58Z"
    analyst: ADMIN
    sample identifier: SMPL2
    equipment serial number: "123456"
  results data:
    glutamine analysis:
      molar concentration: 2.4
      molar concentration unit: mmol/L
    glucose analysis:
      mass concentration: 6.71
      mass concentration unit: g/L
    lactate dehydrogenase analysis:
      molar concentration: BELOW_RANGE
      molar concentration unit: U/L
    ammonia analysis:
      molar concentration: 1.87
      molar concentration unit: mmol/L
    lactate analysis:
      mass concentration: BELOW_RANGE
      mass concentration unit: g/L
```

Highlighting LLM interpretation in our attention tracing tool



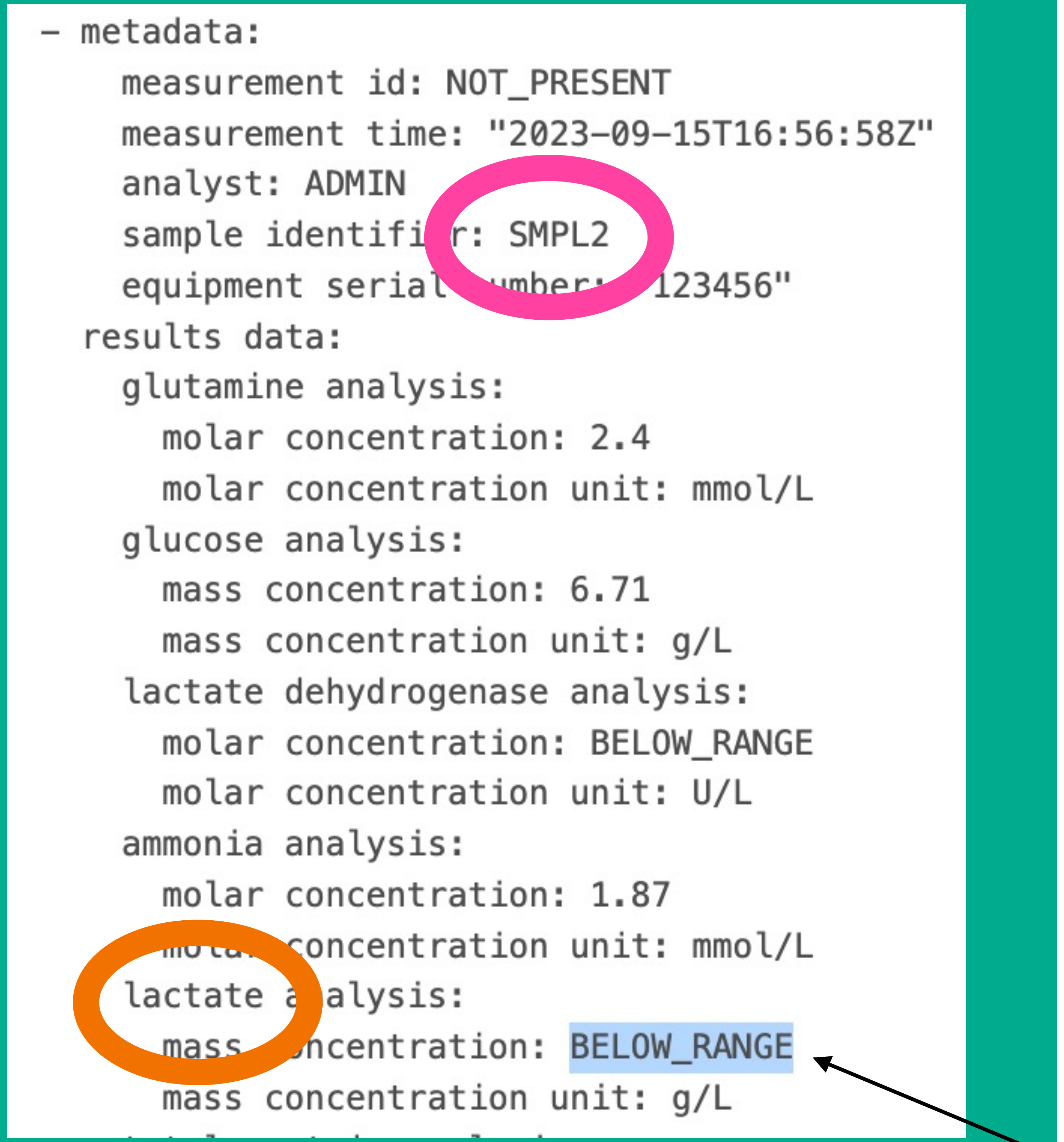
```
0     2023-09-17 13:04:06    #ARC-FILE#    1.1a    2021-05-01    2023-09-17    CEDEX BIO HT    123456  6.0.0.1905 (1905)       ADMIN
40    2023-09-15 16:55:51    SMPL1         SAM                   GLN2B                 mmol/L                2.45     0.17138          R
40    2023-09-15 16:55:53    SMPL1         SAM                   GLC3B                 g/L                   6.32     1.05394          R
40    2023-09-15 16:56:18    SMPL1         SAM                   LDH2B                 U/L                   88.09    0.00728          R
40    2023-09-15 16:56:26    SMPL1         SAM                   NH3B                  mmol/L                1.846    0.05333          R
40    2023-09-15 16:56:37    SMPL1         SAM                   LAC2B                 g/L                   0.02     0.01567          R
40    2023-09-15 16:56:48    SMPL1         SAM                   TP2LB                 g/L                   4.6      0.14883          R
40    2023-09-15 16:56:58    SMPL2         SAM                   GLN2B                 mmol/L                2.40     0.16787          R
40    2023-09-15 16:57:09    SMPL2         SAM                   GLC3B                 g/L                   6.71     1.11766          R
40    2023-09-15 16:57:19    SMPL2         SAM                   LDH2B                 U/L       < TEST RNG      < 20.00    0.00060         R
40    2023-09-15 16:57:30    SMPL2         SAM                   NH3B                  mmol/L                1.870    0.05408          R
40    2023-09-15 16:57:41    SMPL2         SAM                   LAC2B                 g/L       < TEST RNG      < 0.00     0.00310         R
40    2023-09-15 16:57:51    SMPL2         SAM                   TP2B                  g/L       < TEST RNG      < 4.0      0.03322         R
40    2023-09-15 16:58:02    SMPL2         SAM                   TP2D                  g/L       < TEST RNG      < 40.0     0.02653         R
40    2023-09-15 16:58:23    SMPL2         SAM                   TP2LB                 g/L                   4.7      0.15217          R
40    2023-09-15 16:58:34    SMPL3         SAM                   GLN2B                 mmol/L                2.43     0.17049          R
40    2023-09-15 16:58:45    SMPL3         SAM                   GLC3B                 g/L                   6.71     1.11813          R
40    2023-09-15 16:58:55    SMPL3         SAM                   LDH2B                 U/L       < TEST RNG      < 20.00    0.00076         R
```

Raw instrument data

Correctly interprets `< TEST RNG` value as below measurement range

Scalable human validation/introspection via attention tracing

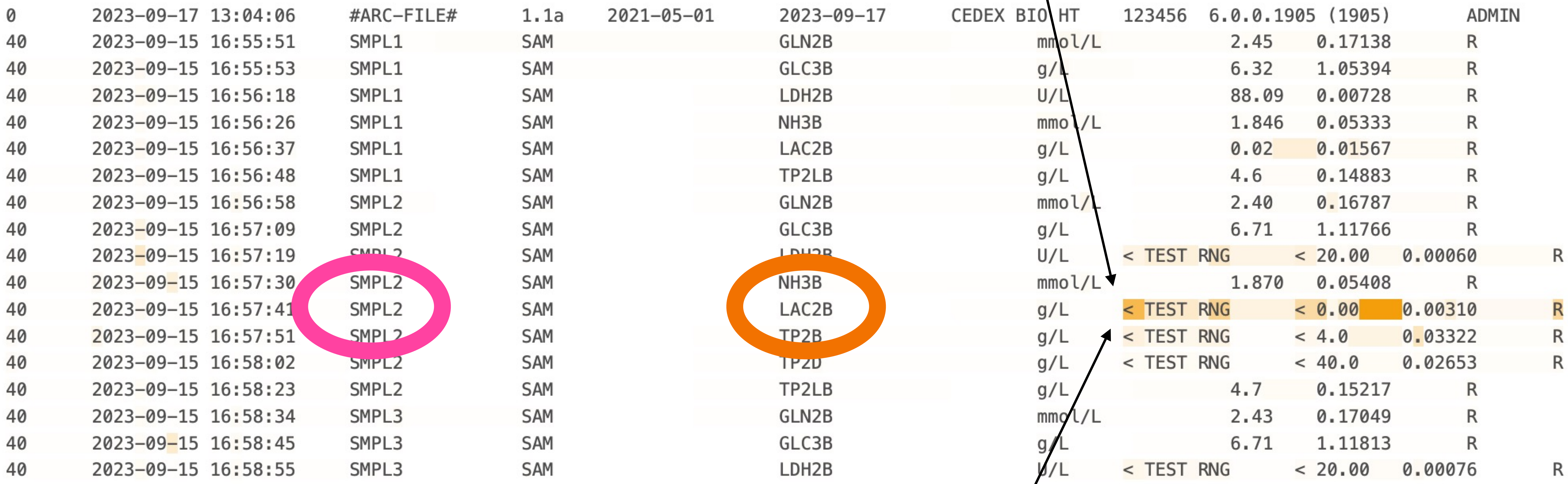
Snapshots from our attention tracing tool



```
– metadata:
    measurement id: NOT_PRESENT
    measurement time: "2023-09-15T16:56:58Z"
    analyst: ADMIN
    sample identifier: SMPL2
    equipment serial number: "123456"
  results data:
    glutamine analysis:
      molar concentration: 2.4
      molar concentration unit: mmol/L
    glucose analysis:
      mass concentration: 6.71
      mass concentration unit: g/L
    lactate dehydrogenase analysis:
      molar concentration: BELOW_RANGE
      molar concentration unit: U/L
    ammonia analysis:
      molar concentration: 1.87
      molar concentration unit: mmol/L
    lactate analysis:
      mass concentration: BELOW_RANGE
      mass concentration unit: g/L
```

Highlighting LLM interpretation in our attention tracing tool

Directs attention to the correct `<  TEST  RNG` value only



```
0    2023-09-17 13:04:06   #ARC-FILE#   1.1a   2021-05-01   2023-09-17   CEDEX BIO HT   123456  6.0.0.1905 (1905)        ADMIN
40   2023-09-15 16:55:51   SMPL1        SAM                 GLN2B                 mmol/L                2.45     0.17138         R
40   2023-09-15 16:55:53   SMPL1        SAM                 GLC3B                 g/L                   6.32     1.05394         R
40   2023-09-15 16:56:18   SMPL1        SAM                 LDH2B                 U/L                   88.09    0.00728         R
40   2023-09-15 16:56:26   SMPL1        SAM                 NH3B                  mmol/L                1.846    0.05333         R
40   2023-09-15 16:56:37   SMPL1        SAM                 LAC2B                 g/L                   0.02     0.01567         R
40   2023-09-15 16:56:48   SMPL1        SAM                 TP2LB                 g/L                   4.6      0.14883         R
40   2023-09-15 16:56:58   SMPL2        SAM                 GLN2B                 mmol/L                2.40     0.16787         R
40   2023-09-15 16:57:09   SMPL2        SAM                 GLC3B                 g/L                   6.71     1.11766         R
40   2023-09-15 16:57:19   SMPL2        SAM                 LDH2B                 U/L     < TEST RNG          < 20.00    0.00060        R
40   2023-09-15 16:57:30   SMPL2        SAM                 NH3B                  mmol/L                1.870    0.05408         R
40   2023-09-15 16:57:41   SMPL2        SAM                 LAC2B                 g/L     < TEST RNG          < 0.00     0.00310        R
40   2023-09-15 16:57:51   SMPL2        SAM                 TP2B                  g/L     < TEST RNG          < 4.0      0.03322        R
40   2023-09-15 16:58:02   SMPL2        SAM                 TP2D                  g/L     < TEST RNG          < 40.0     0.02653        R
40   2023-09-15 16:58:23   SMPL2        SAM                 TP2LB                 g/L                   4.7      0.15217         R
40   2023-09-15 16:58:34   SMPL3        SAM                 GLN2B                 mmol/L                2.43     0.17049         R
40   2023-09-15 16:58:45   SMPL3        SAM                 GLC3B                 g/L                   6.71     1.11813         R
40   2023-09-15 16:58:55   SMPL3        SAM                 LDH2B                 U/L     < TEST RNG          < 20.00    0.00076        R
```

Raw instrument data

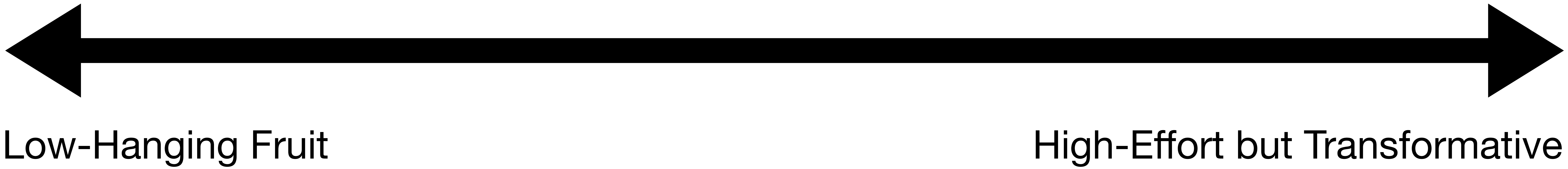
Correctly interprets `<  TEST  RNG` value as below measurement range

We can build converters fast

What else can we do with LLMs to further FAIR data and Allotrope adoption?

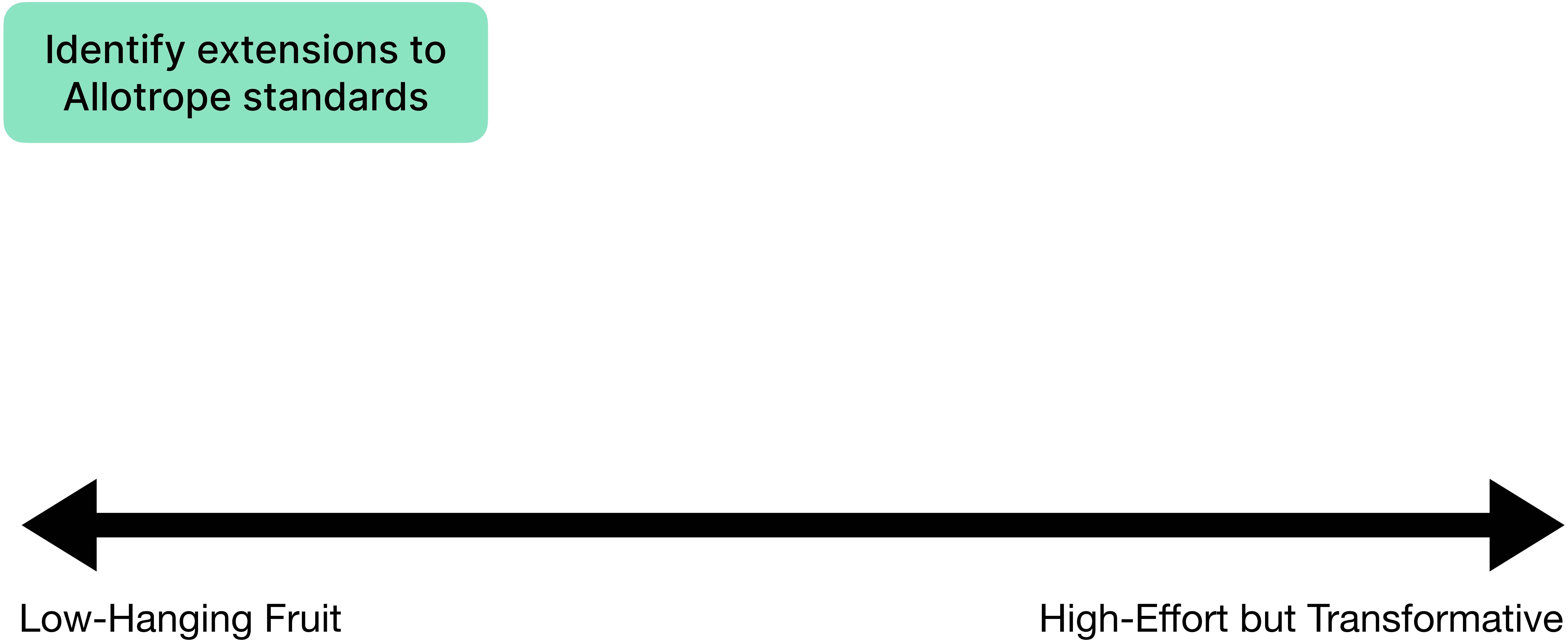
We can build converters fast

What else can we do with LLMs to further FAIR data and Allotrope adoption?



We can build converters fast

What else can we do with LLMs to further FAIR data and Allotrope adoption?



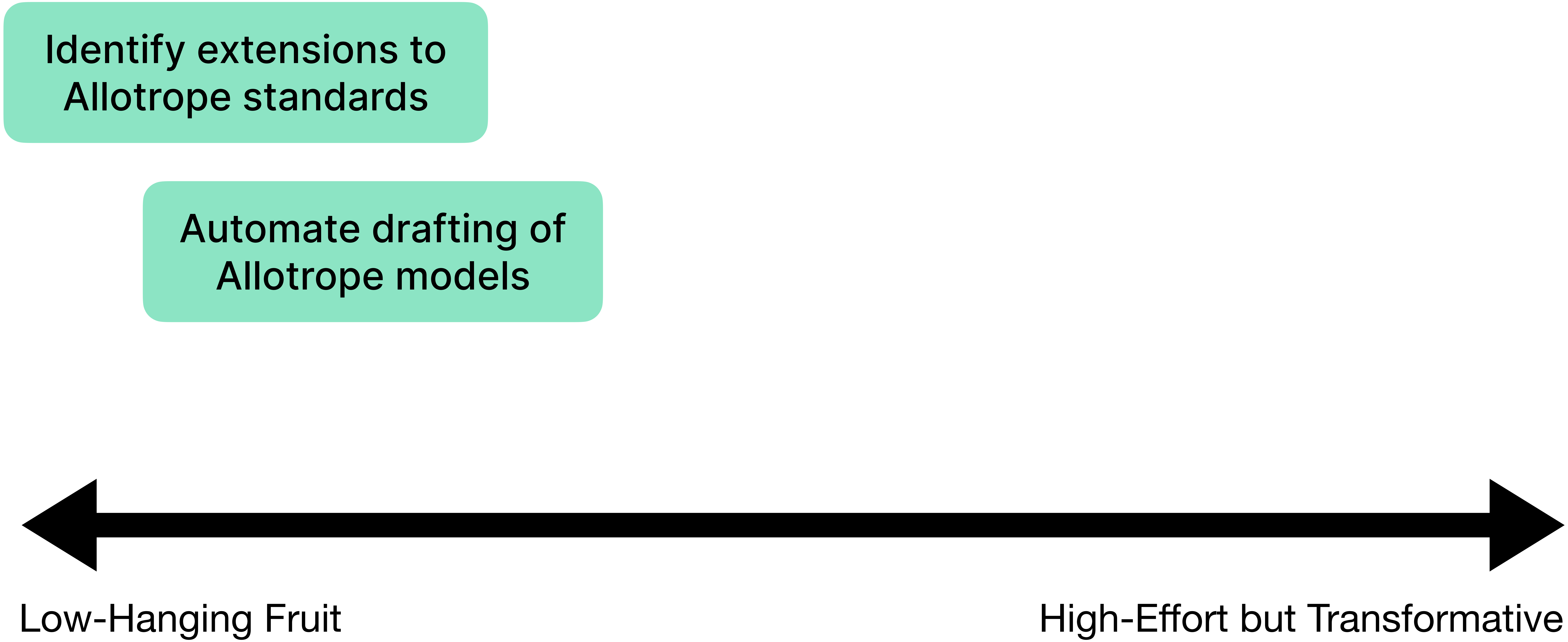
Identify extensions to Allotrope standards

Low-Hanging Fruit

High-Effort but Transformative

We can build converters fast

What else can we do with LLMs to further FAIR data and Allotrope adoption?



Identify extensions to Allotrope standards

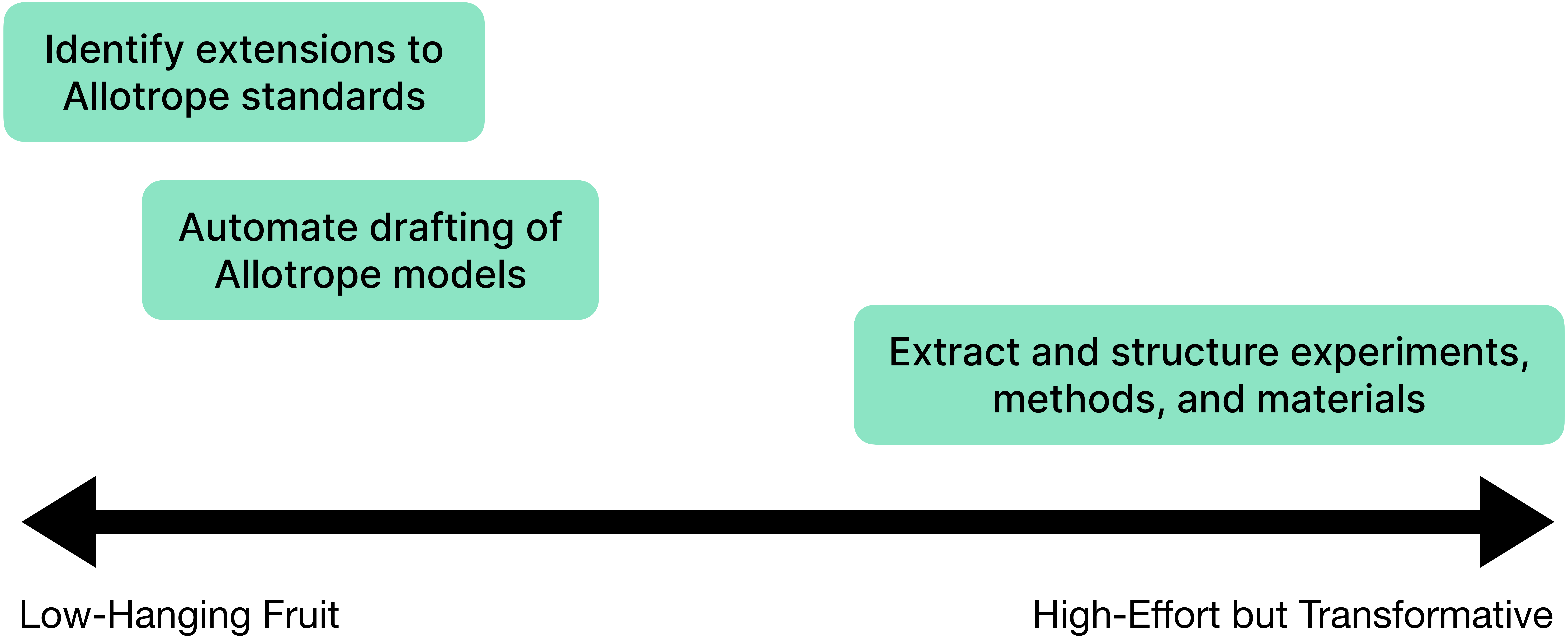
Automate drafting of Allotrope models

Low-Hanging Fruit

High-Effort but Transformative

We can build converters fast

What else can we do with LLMs to further FAIR data and Allotrope adoption?



Identify extensions to Allotrope standards

Automate drafting of Allotrope models

Extract and structure experiments, methods, and materials

Low-Hanging Fruit

High-Effort but Transformative

Thank you!

Questions? Feedback? Comments?

a@awchen.com