

Evaluation pipeline results

python -m evaluation.evaluate --strategy all

Model 1: qwen/qwen3-2b

```
Loaded 10 ground truth labels from evaluation\ground_truth
OCR backend: easyocr | LLM model: qwen/qwen3-32b

Evaluating strategy: regex
Using CPU. Note: This module is much faster with a GPU.
D:\exponent.ch\pdf_ocr\.venv\Lib\site-packages\torch\utils\data\loader.py:775: UserWarning:
nned memory won't be used.
  super().__init__(loader)
[regex] 118NP8.pdf - 33.4s
[regex] cafecepe_20180908_008.pdf - 6.0s
[regex] safeway-3208524524.pdf - 29.9s
[regex] starbucks-20190102.pdf - 5.5s
[regex] monoprix-325435432.pdf - 15.0s
[regex] caffenero_20190831.pdf - 9.8s
[regex] 20210906_001.pdf - 19.7s
[regex] 20181024_008.pdf - 51.7s
[regex] bioburger_20240327_004.pdf - 12.0s
[regex] venetian_434280912998.pdf - 50.2s

=====
Strategy: REGEX | OCR: easyocr | Model: qwen/qwen3-32b
(10 receipts)
=====
Field                Avg Similarity
-----
Provider Name        0.0%
Provider Address     13.4%
VAT Number           75.0%
Currency             80.0%
Total Amount         0.0%
VAT Info             40.0%
Items F1             0.0%
-----
Avg Latency          23.3s
```

```
Evaluating strategy: llm
[llm] 118NP8.pdf - 23.2s
D:\exponent.ch\pdf_ocr\.venv\Lib\site-packages\torch\utils\data\loader.py:7
nned memory won't be used.
  super().__init__(loader)
[llm] cafecepe_20180908_008.pdf - 6.5s
[llm] safeway-3208524524.pdf - 14.5s
[llm] starbucks-20190102.pdf - 9.7s
[llm] monoprix-325435432.pdf - 22.9s
[llm] caffenero_20190831.pdf - 13.9s
[llm] 20210906_001.pdf - 35.4s
[llm] 20181024_008.pdf - 67.8s
[llm] bioburger_20240327_004.pdf - 25.0s
[llm] venetian_434280912998.pdf - 73.0s

=====
Strategy: LLM | OCR: easyocr | Model: qwen/qwen3-32b
(10 receipts)
=====
Field                Avg Similarity
-----
Provider Name        87.0%
Provider Address     71.5%
VAT Number           70.0%
Currency             90.0%
Total Amount         96.2%
VAT Info             11.2%
Items F1             36.0%
-----
Avg Latency          29.2s
```

Model 2: openai/gpt-oss-120b

```
OCR backend: easyocr | LLM model: openai/gpt-oss-120b
Evaluating strategy: hybrid
Using CPU. Note: This module is much faster with a GPU.
D:\exponent.ch\pdf_ocr\.venv\Lib\site-packages\torch\utils\data\data_loader.py:775: UserWarning
  super().__init__(loader)
[hybrid] 118NP8.pdf - 26.4s
D:\exponent.ch\pdf_ocr\.venv\Lib\site-packages\torch\utils\data\data_loader.py:775: UserWarning
  super().__init__(loader)
[hybrid] cafecepe_20180908_008.pdf - 7.5s
[hybrid] safeway-3208524524.pdf - 16.8s
[hybrid] starbucks-20190102.pdf - 8.0s
[hybrid] monoprix-325435432.pdf - 13.2s
[hybrid] caffenero_20190831.pdf - 8.3s
[hybrid] 20210906_001.pdf - 16.9s
[hybrid] 20181024_008.pdf - 49.6s
[hybrid] bioburger_20240327_004.pdf - 14.1s
[hybrid] venetian_434280912998.pdf - 58.1s

=====
Strategy: HYBRID | OCR: easyocr | Model: openai/gpt-oss-120b
(10 receipts)
=====
Field                Avg Similarity
-----
Provider Name        92.6%
Provider Address     70.8%
VAT Number           70.8%
Currency              90.6%
Total Amount         100.6%
VAT Info              6.6%
Items F1             59.7%
-----
Avg Latency          21.9s
(pdf_ocr)\BF-Polymorph-011.pdf - 0
```