ConvFinQA Model Evaluation Summary

Training Data Construction

The training data was built from the `train.json` file in the ConvFinQA dataset using a preprocessing script. The process included:

- Extracting question-answer pairs from both `qa` and indexed fields like `qa_0`, `qa_1`, etc.
- Filtering out questions without answers.
- Normalising questions (lowercasing, punctuation removal).
- Creating a lookup dictionary (`lookup_qa`) to pair each question with its answer.
- Augmenting important financial questions with phrasing variations to improve generalisation.
- The final dataset was split to ensure the evaluation set (`lookup_finetune_test.json`) was held out from training.

Model and Training Summary

I trained a `google/flan-t5-base` model using a memorisation-style fine-tuning pipeline:

- Input Format: `Question: <text>`
- Target Format: `Answer: <text>`
- Model Type: `AutoModelForSeq2SeqLM` from Hugging Face Transformers
- Training handled via `Seq2SeqTrainer`
- The model learns direct mappings between financial QA pairs, relying on fine-tuned textual reasoning.

Model Performance

The trained model was evaluated on 1,029 held-out QA examples.

Overall Metrics

Metric	Value
Total Examples	1,029
Exact Match Accuracy	89.31%
Partial Match Accuracy	89.41%
F1 Score (Weighted)	0.8914
Precision (Weighted)	0.9054
Recall (Weighted)	0.8931

Answer Type Breakdown

Answer Type	Accuracy	Correct / Total
Percentage	89.4%	707 / 791
Numeric	88.7%	196 / 221
Currency	100.0%	10 / 10
Text	100.0%	7 / 7

Findings

- The model generalises well on percentage-based financial questions, which are the most common in the dataset.
- Currency and numeric responses are also handled with strong accuracy.
- All text-based answers were correct, though they were few in number.
- There is a high overlap between exact and partial match scores, indicating that the model produces structurally correct and semantically equivalent answers consistently.
- The F1, precision, and recall scores reflect a strong balance of correctness and completeness.

Conclusion

This evaluation demonstrates that the model is capable of accurately answering short-form financial questions when trained with a memorisation-style pipeline. It benefits from data augmentation and question normalisation. Although it does not reason over tables directly, it achieves strong performance by learning robust mappings from questions to answers.

Future Scope

If the dataset were to shift towards formats that include pre- and post-text with tabular data rather than direct QA pairs, a more agentic architecture would be required. This would likely involve:

- Retrieval of relevant financial statements or text segments
- Numerical reasoning via a calculator tool
- Possibly incorporating multi-hop reasoning or table parsing capabilities

This would transition the system from pure memorisation to contextual reasoning.