Domain-adaptive Post-training For Financial LLMs

Zixuan Ke

https://vincent950129.github.io/

FinDAP

HIghlighted outcomes

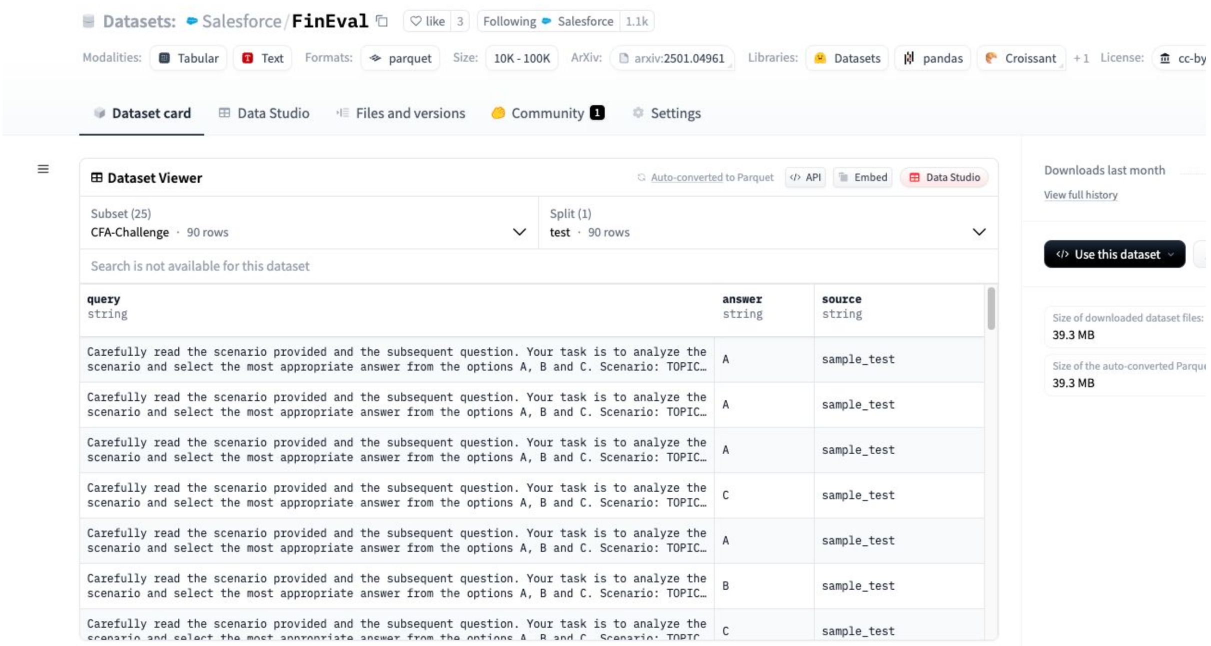
Resulting model (Llama-Fin-8b), a small but mighty Finance LLM!

Benchmark	Llama-Fin (Ours, 8b)	GPT4o
FPB (Financial Sentiment analysis)	91.13	82.16
FiQA SA (Financial Sentiment analysis)	95.32	68.51
NER (Financial Named Entity Recognition)	76.69	43.02
EDTSUM (Financial Abstractive Summarization)	53.78	18.15
Finance Bench (Financial Open QA)	54.00	51.30
SM-Bigdata (Stock Movement Prediction)	54.14	49.18
Flare-German (Credit Scoring)	64.00	17.00
……		

FinDAP

Outcome

Evaluation set is open source



https://huggingface.co/datasets/Salesforce/FinEval

Benchmark	Llama-Fin (Ours, 8b)	GPT4o
FPB (Financial Sentiment analysis)	91.13	82.16
FiQA SA (Financial Sentiment analysis)	95.32	68.51
NER (Financial Named Entity Recognition)	76.69	43.02
EDTSUM (Financial Abstractive Summarization)	53.78	18.15
Finance Bench (Financial Open QA)	54.00	51.30
SM-Bigdata (Stock Movement Prediction)	54.14	49.18
Flare-German (Credit Scoring)	64.00	17.00
……		

Research Question

Given a strong general-purpose LLM (*e.g.*, Llama3-8b-inst), how to effectively adapt it to a target domain (*e.g.*, finance) by post-training? What criteria are desirable for successful adaptation? What are effective training recipes with respect to data and model?

Research Question

Answer from related work:
(e.g., PIXIU, FinLLM, FinTral, Palmyra-Fin, FinMa, Finance-LLM, FinLLaVA)*

Follow standard methods:

Continual Pre-training (CPT) → Instruction-tuning (IT) → Preference Alignment (PA)

Given a strong general-purpose LLM (*e.g.*, Llama3-8b-inst), how to effectively adapt it to a target domain (*e.g.*, finance) by post-training? What criteria are desirable for successful adaptation? What are effective training recipes with respect to data and model?

Finance LLM	Capabilities	Recipe		Evaluation
		Model Recipe	Data Recipe	
AdaptLLM	Concept	CPT	**CPT**: Financial text + heuristic QAs constructed from the text	Financial + Classification tasks + Direct answer
PIXIU	Task	IT	**IT**: Financial tasks	Financial + Classification tasks + Direct answer
FinLLM	Concept, Task	CPT → IT	**CPT**: Financial text + Fineweb; **IT**: Filtered Financial tasks	Financial + Classification tasks + Direct answer
FinTral	Concept, Task	CPT → IT → PA	**CPT**: Financial text; **IT**: Financial tasks; **PA**: Outcome signal only	Financial + Classification tasks + Direct answer
Palmyra-Fin	SoTA public checkpoint, but recipe is not disclosed			

Research Question

Answer from related work:
(e.g., PIXIU, FinLLM, FinTral, Palmyra-Fin etc.)*

Follow standard methods:

Continual Pre-training (CPT) → Instruction-tuning (IT) → Preference Alignment (PA)

This work (FinDAP)

This is not enough! Domain-adaptive post-training is unique to pre-training and general post-training and we need a systematic and principle approach

Given a strong general-purpose LLM (*e.g.*, Llama3-8b-inst), how to effectively adapt it to a target domain (*e.g.*, finance) by post-training? What criteria are desirable for successful adaptation? What are effective training recipes with respect to data and model?

FinDAP

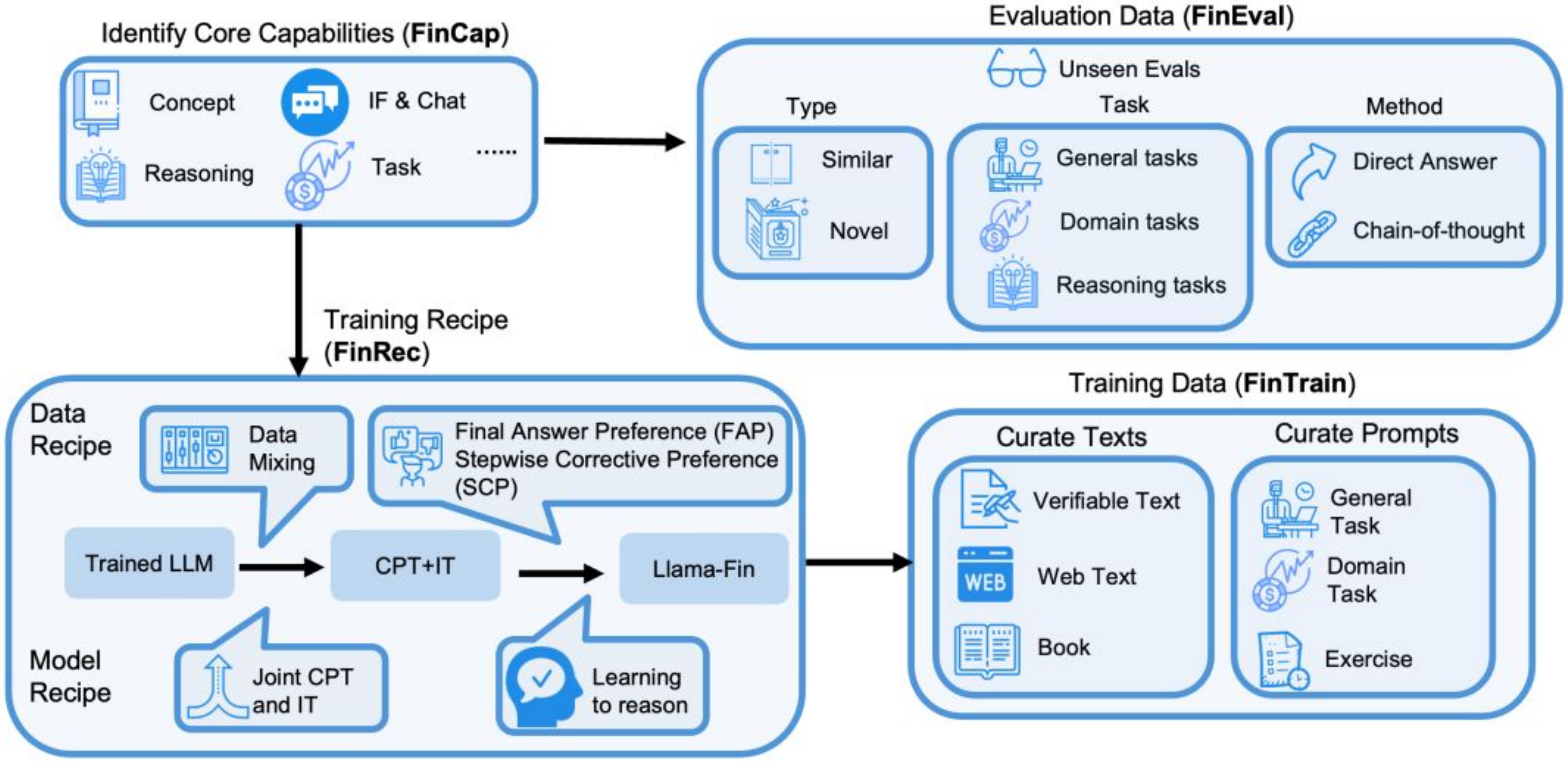
Factors to consider

- **Desirable capabilities** for the target domain (e.g., reasoning...)
- **Training Recipe**
  - Original pre-trained LLM already possess strong general capabilities and knowledge
    - Catastrophic Forgetting
    - Knowledge Transfer
  - Construct preference data for reasoning in preference alignment (PA)
- **Implementation of the recipe** (training data)
  - Quantity vs. Quality
    - Literature found that small amount of general data is enough to mitigate forgetting
    - While learning domain-specific knowledge typical require more data
- **Evaluation**
  - Different capabilities may require different evaluation methods
    - e.g., reasoning tasks may want CoT evaluation

FinDAP

Our Framework

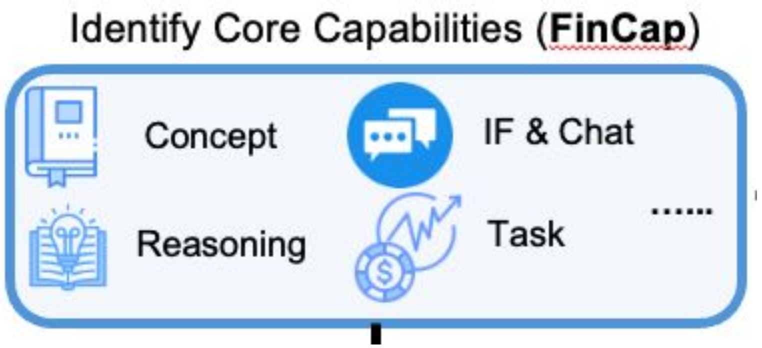
- **FinCap:** Core capacities required for finance domain
- **FinRec:** Our training Recipe
- **FinTrain**: a curated set of training datasets implement FinRec
- **FinEval:** A comprehensive evaluation framework



FinDAP

FinCap: Core capacities required for finance domain

- We propose 4 main capabilities based on the fundamental requirement in FinAI
  - Understanding **domain-specific concepts** to process financial language accurately, performing **domain-specific tasks** to solve real-world problems, **reasoning** effectively to analyze complex financial data, and **following instructions** to interact naturally in practical applications.
  -

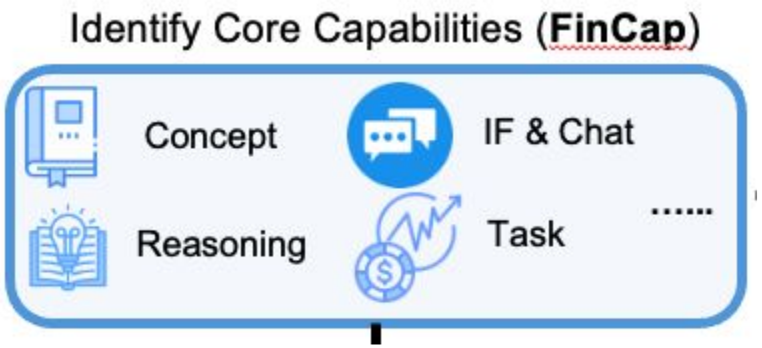


- **Domain-specific concepts**
- **Domain-specific tasks**
- **Instruction-following**
- **Reasoning**

FinDAP

FinCap: Core capacities required for finance domain

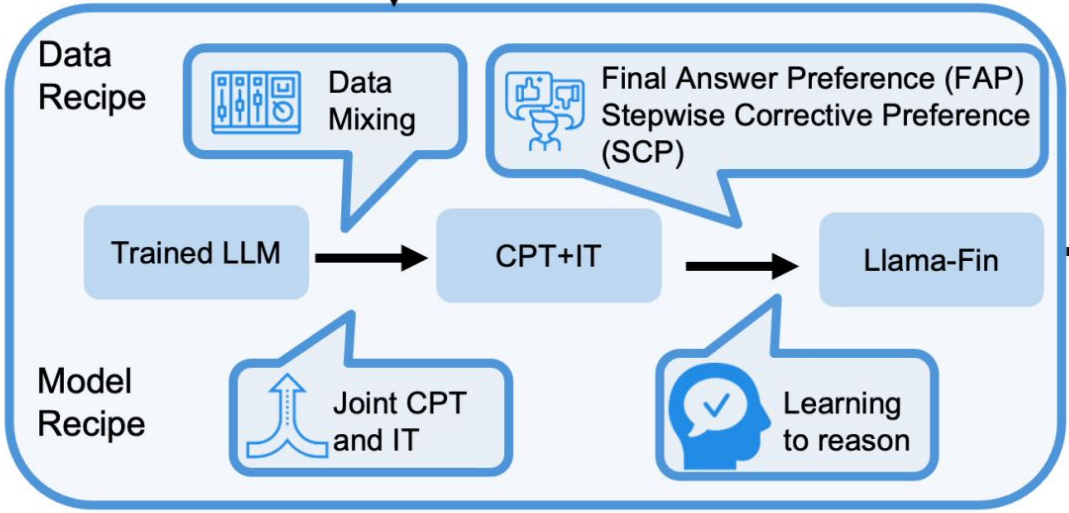
- We propose 4 main capabilities based on the fundamental requirement in FinAI
  - Understanding **domain-specific concepts** to process financial language accurately, performing **domain-specific tasks** to solve real-world problems, **reasoning** effectively to analyze complex financial data, and **following instructions** to interact naturally in practical applications.
  -



- **Domain-specific concepts** → Continual Pre-training (CPT)
- **Domain-specific tasks** → Instruction Tuning (IT)
- **Instruction-following** → Instruction Tuning (IT)
- **Reasoning** → Preference Alignment (PA)

FinDAP

FinRec: Our training Recipe

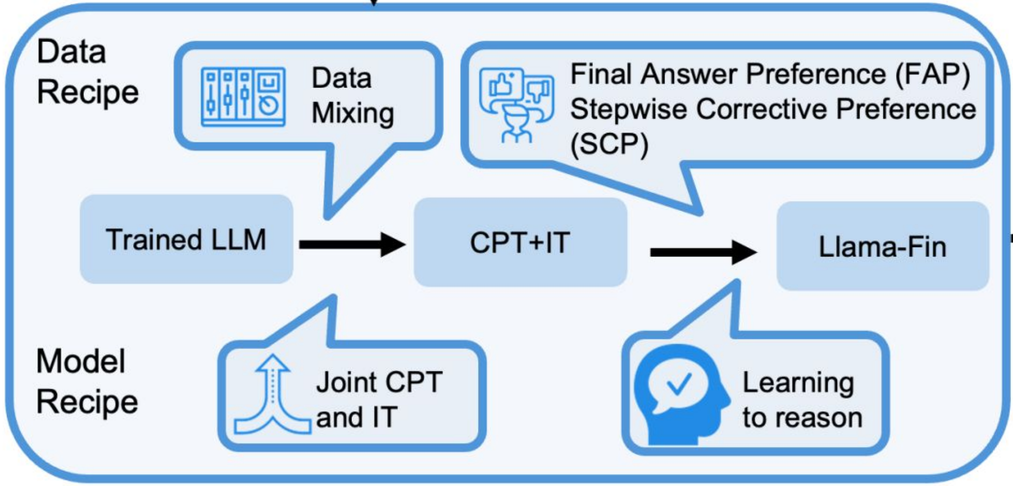


- Model Recipe
  - **Joint training CPT and IT**
    - **Why?**
      - **CPT alone causes forgetting on instruction-following abilities.**
      - **A joint training can further improve generalization**
        - Concepts are often inherently more generalizable due to the shared nature of concepts across tasks
      - **Implementation**
        - Since the only different is whether to mask-out the instruction, we can simply mixing up their data to achieve jointly training
        - CPT data size is usually larger, we down-sample it to match the IT size
  - **PA for reasoning tasks**
    - Assign higher probability mass to better generations, has been shown to be effective in enhancing reasoning capabilities of LLMs
    - Employ DPO (detailed in data recipe)

*All claims are supported by ablation results, which we did not include here (see Appendix)

FinDAP

FinRec: Our training Recipe



- Data Recipe
  - **In-domain, general-domain and mixture**
    - **Most FinLLMs use in-domain data only**
      - This exclusive reliance on in-domain data can lead to forgetting of general knowledge in the original pre-trained LLM.
    - **We conduct systematic investigation ({CPT/IT/PA}-{In/Gen/Mix})**
      - **CPT**
        - While CPT-In and CPT-Gen outperforms in financial and general tasks, respectively, CPT-Mix achieves the best → mixing data sources effectively mitigates forgetting of general knowledge
      - **IT**
        - IT-Mix slightly outperforms than other data versions → mixing general tasks remains helpful to mitigating forgetting of general concepts and tasks, although the effect is much less pronounced compared to CPT.
      - **PA.**
        - PA-In performs comparably to PA-Mix, indicating that it is NOT essential to include general tasks to prevent forgetting of concepts or tasks, unlike the cases of CPT and IT.

Forgetting impact decrease

Mixture of in- and general-domain data for CPT+IT

*All claims are supported by ablation results, which we did not include here (see Appendix)

FinDAP

FinRec: Our training Recipe

- Data Recipe
  - **Preference data with outcome and process signals/reward**
    - **Two Regimes**: Learning to reason (DS-R1…) vs inference scaling (OAI-O1..)
      - We adopt "**learning to reason**" as finance domain often require rapid responses
    - **Learning to reason**
      - Trajectories collection
        - Search-based
          - RM/verifier to guide the search
        - Revision-based
          - Iterative refinement
        - We adopt the **search-based method** as revision-based shows mixed results and have not yet been well established as reliable for achieving improvements
      - Training from trajectories (DPO in this work)

*All claims are supported by ablation results, which we did not include here (see Appendix)

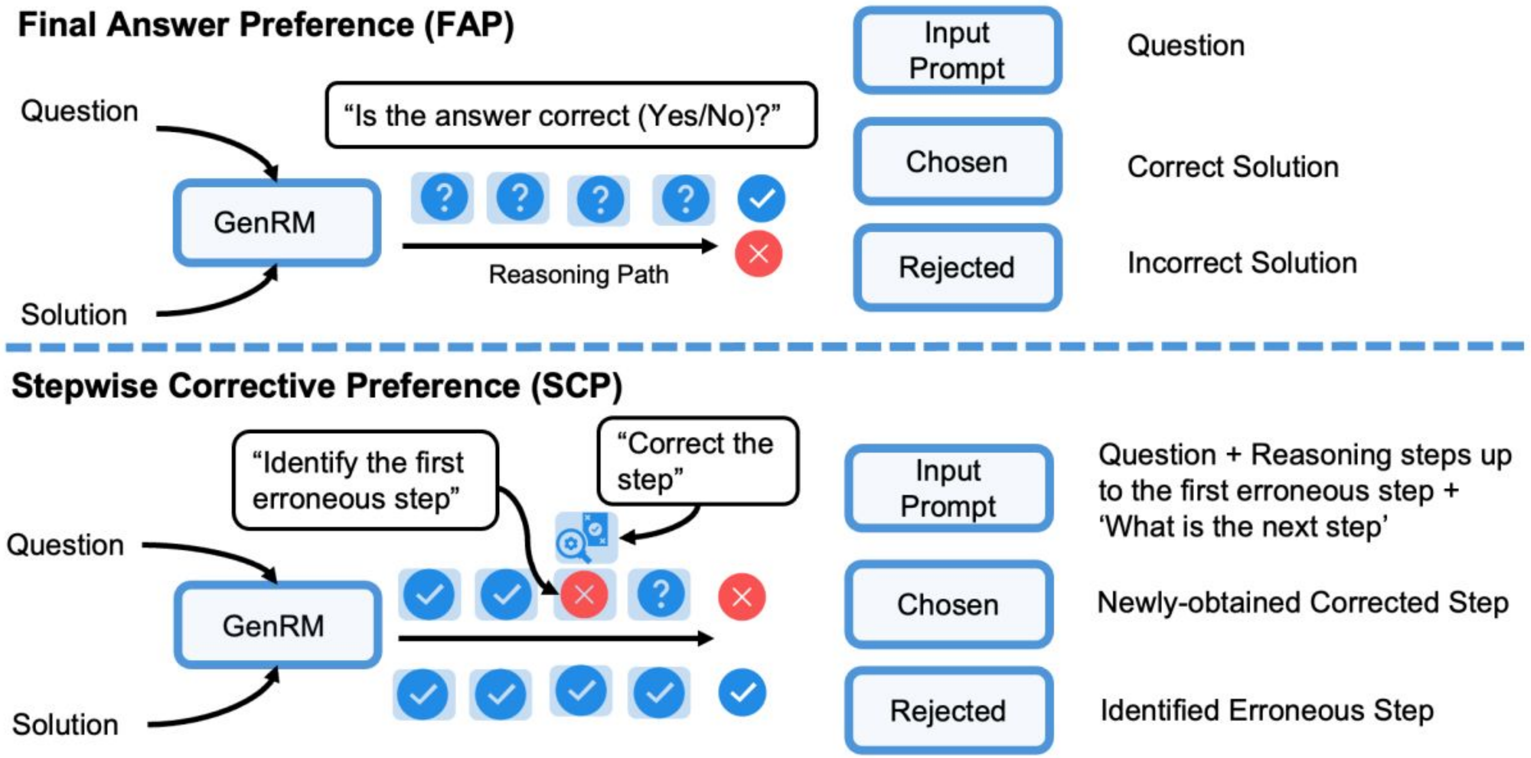
FinDAP

FinRec: Our training Recipe

- Learning to reason: Search-based trajectories collection
  - Reward model / verifier
    - We employ generative RM with strong pre-trained LLM (i.e., GPT4o)

Final answer
(Outcome) Level

Reasoning steps
(Process) Level

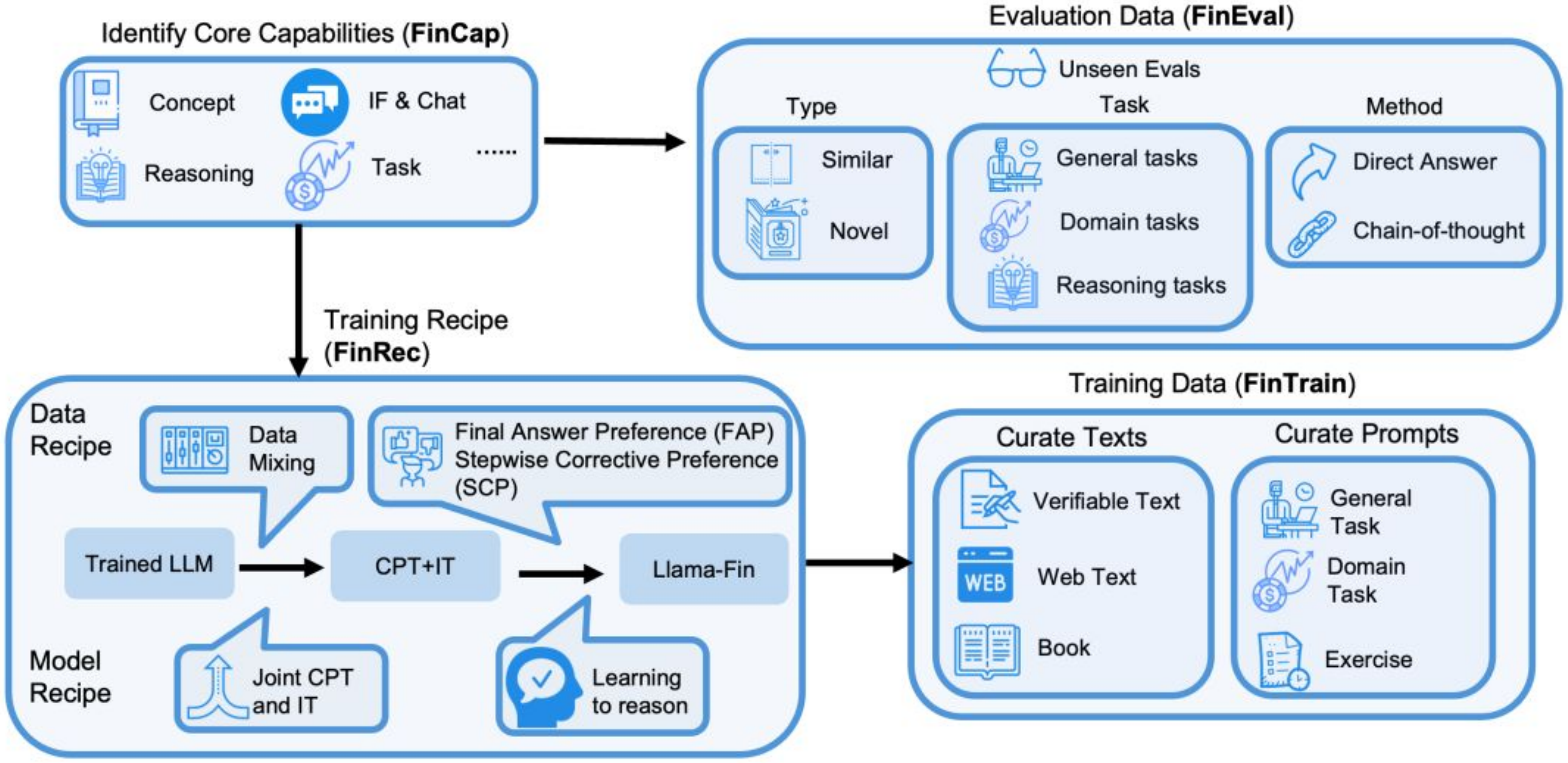


*All claims are supported by ablation results, which we did not include here (see Appendix)

FinDAP

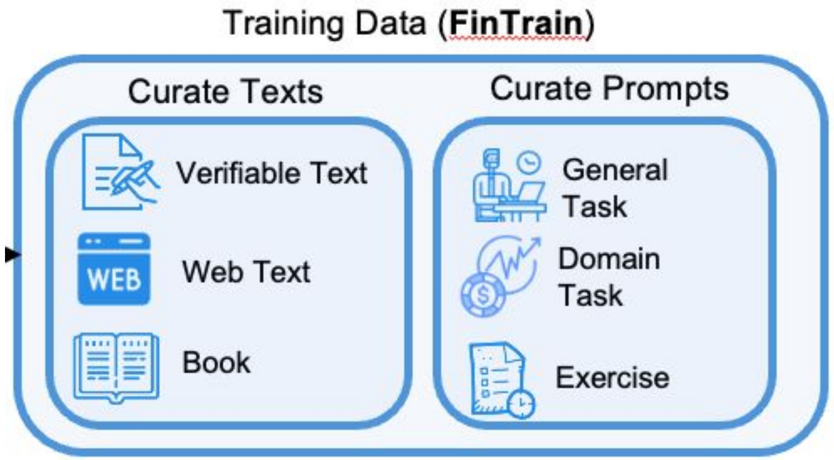
Our Framework

- **FinCap:** Core capacities required for finance domain
- **FinRec:** Our training Recipe
- **FinTrain**: a curated set of training datasets implement FinRec
- **FinEval:** A comprehensive evaluation framework



FinDAP

FinTrain: a curated set of training datasets implement FinRec



- **General text (used in CPT)**
  - **Goal**: mitigate forgetting
  - **Literature**: a 'small' amount of general text (as little as 1%) can effectively mitigate the forgetting issue
  - **FinDAP**: focus on collecting a relatively small but high-quality set of general-domain text.
  - **Verifiable:** text written by humans and previously used in supervised task
- **Finance text (used in CPT)**
  - **Goal:** domain-specific knowledge
  - Diverse and large-scale:
    - Web (URLs based filtering)
    - Books

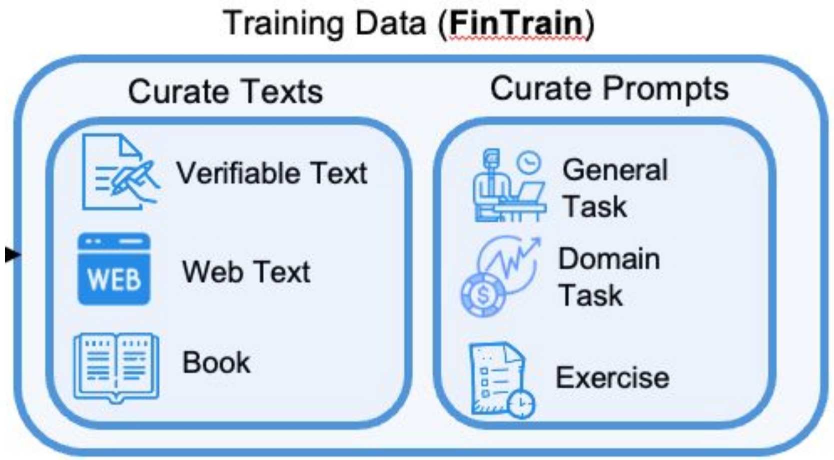
CPT datasets totally ~ 6B tokens

Capability	Domain	CPT Dataset	Size	Reference
Concept	General	NaturalInstrution	100,000	Mishra et al. (2022)
		PromptSource	100,000	Bach et al. (2022)
		Math	29,837	Amini et al. (2019b)
		Aqua	97,500	Ling et al. (2017)
		CREAK	10,200	Onoe et al. (2021)
		ESNLI	549,367	Camburu et al. (2018)
		QASC	8,130	Khot et al. (2020)
		SODA	1,190,000	Kim et al. (2022)
		StrategyQA	2,290	Geva et al. (2021)
		UnifiedSKG	779,000	Xie et al. (2022)
		GSM8K	7,470	Cobbe et al. (2021)
		ApexInstr	1,470,000	Huang et al. (2024b)
		DeepmindMath	379,000	Saxton (2019)
		DialogueStudio	1,070,000	Zhang et al. (2023)
	Finance	**Fineweb-Fin**	4,380,000	-
		Book-Fin	4,500	-
Total			10,177,294	

Table 3: Summary of curated texts. New datasets released with FinDAP are color-highlighted for emphasis.

FinDAP

FinTrain: a curated set of training datasets implement FinRec



- **Prompts (used IT and PA)**
  - Corresponding to each capabilities
  - Diversity
  - Previously shown perwell well (e.g., UltraQA)
  - Potential reasoning steps provided (e.g., Exercise)

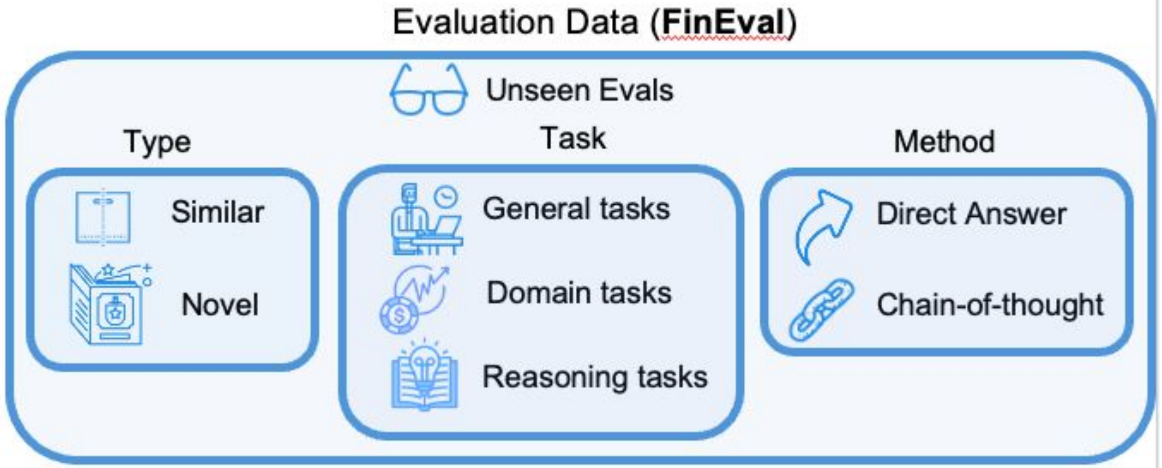
Capability	Domain	Task	IT Dataset	Size	Reference
Tasks	Finance	Relation Cls.	FingptFinred	27,600	Sharma et al. (2022)
		NER	FingptNERCls	13,500	Yang et al. (2023)
			FingptNER	511	Alvarado et al. (2015)
		Headline Cls.	FingptHeadline	82,200	Sinha et al. (2020)
		Sentiment Cls.	SentimentCls	47,600	Yang et al. (2023)
			SentimentTra	76,800	Yang et al. (2023)
		Summariz.	TradeTheEvent	258,000	Zhou et al. (2021)
IF/Chat	General	IF/Chat	SelfInstruct	82,000	Wang et al. (2022)
			SlimOrca	518,000	Lian et al. (2023)
			UltraChat	774,000	Ding et al. (2023)
			ShareGPT	100,000	Link
	Finance	QA	FinanceInstruct	178,000	Link
			FingptConvfinqa	8,890	Chen et al. (2022)
			FlareFinqa	6,250	Chen et al. (2021)
			FlareFiqa	17,100	Yang et al. (2023)
Reasoning	Math	QA	OrcaMath	200,000	Mitra et al. (2024)
			MetaMathQA	395000	Yu et al. (2023)
			MathInstruct	262,000	Xiang Yue (2023)
	Code	QA	MagicodeInstruct	111,000	Luo et al. (2023)
	Finance	CFA Exam	**Exercise**	2,950	-
Total				3,161,401	

Table 4: Summary of our curated prompts. New datasets released with FinDAP are color-highlighted for emphasis. For datasets without formal references but only a URL, we provide their links.

FinDAP

FinEval: A comprehensive evaluation framework

- **Types**
  - Similar (to training)
    - The task type has been seen
    - **Goal**: outperform the best model (GPT4o..), even if we are small (post-training from LLama3-8b-instruct)
  - Novel
    - A new task type
    - **Goal**: outperform our original pre-train LLM (LLama3-8b-instruct)
- **Tasks**
  - Corresponding to the 4 capabilities
- **Methods**
  - CoT for reasoning tasks



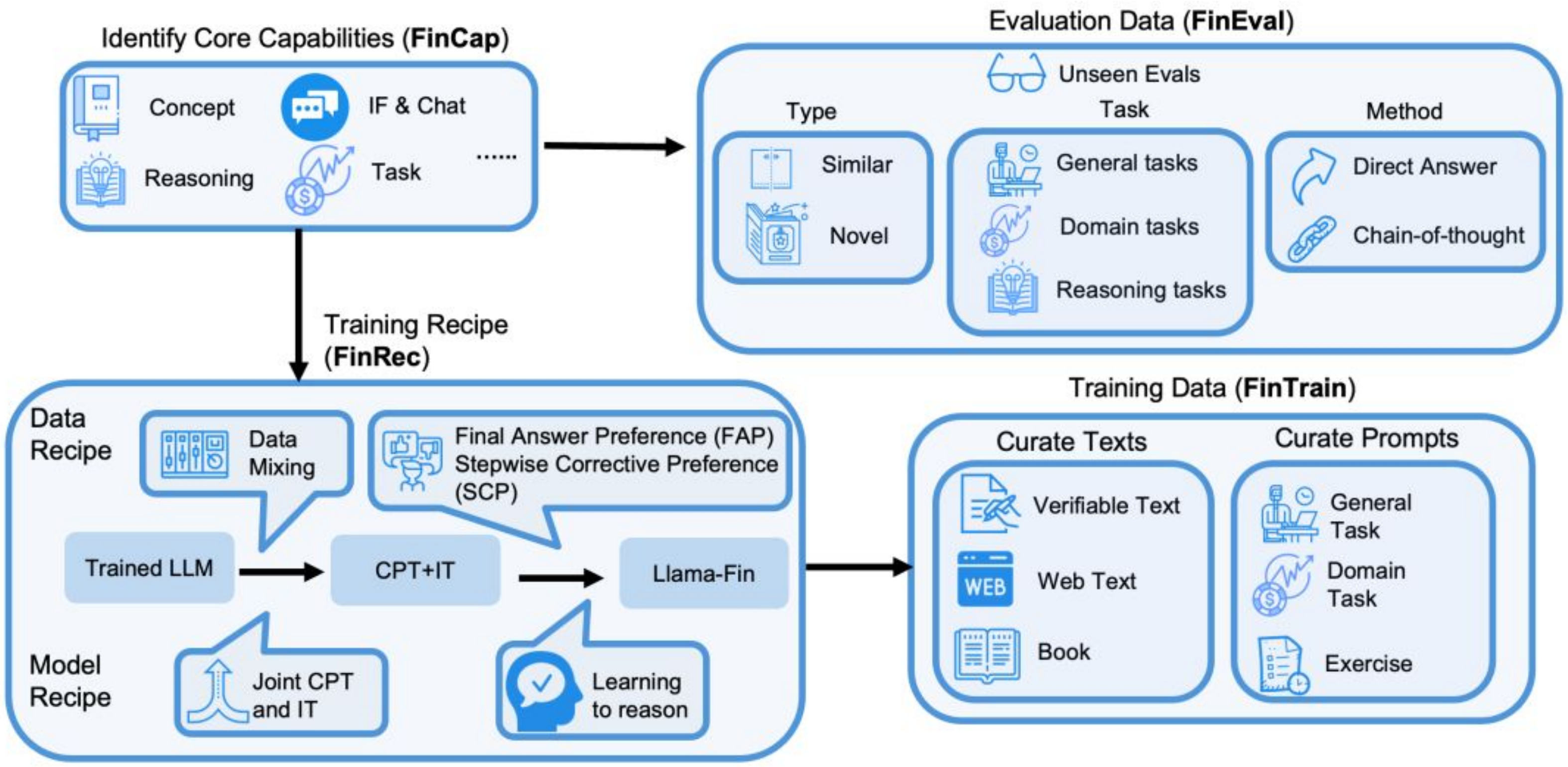
Capability	Domain	Task	Evaluation Dataset	Size	Reference
			Unseen - Similar		
Tasks	Finance	Sentiment Analysis	FPB	970	Malo et al. (2014)
			FiQA SA	235	Maia et al. (2018)
		Monetary policy Stance	FOMC	496	Shah et al. (2023)
		Named entity recognition	NER	98	Alvarado et al. (2015)
		Abstractive Summarization	EDTSUM	2,000	Zhou et al. (2021)
Total				3,799	
			Unseen - Novel		
Concept	General	Knowledge Recall	MMLU	14,042	(Hendrycks et al., 2021)
			AI2-ARC	3,548	Clark et al. (2018)
			Nq-open	7,842	Kwiatkowski et al. (2019)
	Finance		**MMLU-Finance**	1,460	-
Tasks	Finance	Extractive Summarization	Flare-ECTSUM	495	Mukherjee et al. (2022)
		ESG Issue Classification	MLESG	300	Chen et al. (2023b)
		Rumour Detection	MA	500	Yang et al. (2020)
		Stock Movement Prediction	SM-Bigdata	1,470	Soun et al. (2022)
			SM-ACL	3,720	Xu and Cohen (2018)
			SM-CIKM	1,140	Wu et al. (2018)
		Fraud Detection	CRA-CCF	2,280	Feng et al. (2024)
			CRA-CCFraud	2,100	Feng et al. (2024)
		Credit Scoring	Flare-German	200	Hofmann (1994)
			Flare-Astralian	139	Quinlan (1987)
			CRA-LendingClub	2,690	Feng et al. (2024)
		Distress Identification	CRA-Polish	1,740	Feng et al. (2024)
			CRA-Taiwan	1,370	Feng et al. (2024)
		Claim Analysis	CRA-ProroSeguro	2,380	Feng et al. (2024)
			CRA-TravelInsurance	2,530	Feng et al. (2024)
		Tabular QA	Flare-TATQA	1,670	Zhu et al. (2021)
		Open QA	Finance Bench	150	Islam et al. (2023)
IF/Chat	General	Precise IF	MT-bench	80	Zheng et al. (2023)
Reasoning	Math	Reasoning	MathQA	2,985	Amini et al. (2019a)
	General	Social Reasoning	Social-IQA	2,636	Welbl et al. (2017)
		Common Reasoning	Open-book-qa	500	Mihaylov et al. (2018)
			Hellaswag	10,003	Zellers et al. (2019)
			Winogrande	1,767	Sakaguchi et al. (2019)
			PIQA	3,000	Bisk et al. (2020)
	Finance	Exam	CFA-Easy	1,030	Link
			CFA-Challenge	90	-
Total				91,872	

Table 1: Summary of our evaluation dataset. New datasets released with FinDAP are color-highlighted for emphasis.

FinDAP

Our Framework

- **FinCap:** Core capacities required for finance domain
- **FinRec:** Our training Recipe
- **FinTrain**: a curated set of training datasets implement FinRec
- **FinEval:** A comprehensive evaluation framework



FinDAP

Our Framework

Finance LLM	Capabilities	Recipe		Evaluation
		Model Recipe	Data Recipe	
AdaptLLM	Concept	CPT	**CPT**: Financial text + heuristic QAs constructed from the text	Financial + Classification tasks + Direct answer
PIXIU	Task	IT	**IT**: Financial tasks	Financial + Classification tasks + Direct answer
FinLLM	Concept, Task	CPT $\rightarrow$ IT	**CPT**: Financial text + Fineweb; **IT**: Filtered Financial tasks	Financial + Classification tasks + Direct answer
FinTral	Concept, Task	CPT $\rightarrow$ IT $\rightarrow$ PA	**CPT**: Financial text; **IT**: Financial tasks; **PA**: Outcome signal only	Financial + Classification tasks + Direct answer
Palmyra-Fin	SoTA public checkpoint, but recipe is not disclosed			
Llama-Fin	Concept, IF/Chat, Task, Reasoning	CPT+IT $\rightarrow$ PA	**CPT**: Financial + General text. **IT**: Financial + General tasks **PA**: A novel PA that leverages outcome and process signals	General + Financial tasks; Similar + Novel tasks Classification + Open-form QA tasks Knowledge Recall + Reasoning tasks Direct answer + CoT

Table 1: Comparison between Llama-Fin with other finance LLMs.

FinDAP

Final Results (similar tasks)

Outperforms all other baselines (including GPT4o) with one exception

Task	Benchmark	Llama-Fin 8B	Llama3 Instruct 8B	Llama3.1 Instruct 8B	Palmyra Fin 70B	Phi 3.5-mini Instruct 3.8B	Mistral Nemo instruct 12B	GPT4o
Sentiment Ana.	FPB (Acc)	**91.13**✓	73.09	71.55	67.11	78.04	78.25	82.16
Sentiment Ana.	FiQA SA (Acc)	**95.32**✓	77.87	70.64	71.91	69.36	55.74	68.51
Monetary Policy	FOMC (Acc)	**64.31**✓	56.65	54.64	63.10	58.47	57.86	67.94
Named Entity	NER (Rouge1)	**76.69**✓	45.03	51.22	54.29	39.37	49.84	43.02
Abs Summ.	EDTSUM (Rouge1)	**53.78**✓	11.50	12.53	21.77	19.97	12.32	18.15

Table 2: Results on **similar (unseen)** tasks. '*' indicates that 'GPT4o' is used as the judge. Llama-Fin and its variant without PA (i.e., the 'CPT+IT' checkpoint) are highlighted in blue while the closed model is highlighted in gray. The best performing model for 8b on each benchmark is **bolded**. The overall best performance across all models is underlined. ✓ indicates that Llama-Fin outperforms the base Llama3-8b-inst.

FinDAP

Final Results (novel tasks)

General concepts are preserved

Effective in the majority of Tasks (12/17)

Instruction-following is also preserved

Excels in reasoning tasks

Capability	Domain	Task	Benchmark	Llama-Fin 8B	Llama3 Instruct 8B	Llama3.1 Instruct 8B	Palmyra Fin 70B	Phi 3.5-mini Instruct 3.8B	Mistral Nemo instruct 12B	GPT4o
Concept	General	Knowledge Recall	MMLU (CoT, Acc)	47.42	**48.14**	47.42	54.93	45.07	49.64	63.88
			AI2-ARC (CoT, Acc)	89.43✓	89.29	**89.80**	89.01	87.25	88.19	97.85
			Nq-open (CoT, Acc)	19.20✓	18.47	**22.52**	19.25	6.20	17.01	27.92
	Finance	Knowledge Recall	MMLU-Finance (Acc)	64.20	65.71	**66.74**	75.15	68.17	61.88	86.52
Task	Finance	Extractive Summ.	Flare-ECTSUM (Rouge1)	34.10	**35.92**	35.77	33.24	35.52	37.86	35.90
		ESG Issue	MLESG (Acc)	**40.67**✓	36.33	36.00	39.67	38.33	32.67	45.67
		Rumor Detection	MA (Acc)	84.00✓	82.60	**84.20**	62.60	75.40	85.20	73.80
		Stock Movement	SM-Bigdata (CoT, Acc)	54.14	**55.3**	46.06	48.70	53.26	53.53	49.18
			SM-ACL (CoT, Acc)	**51.99**✓	50.51	45.30	51.21	49.84	50.75	50.97
			SM-CIKM (CoT, Acc)	54.94	**55.56**	48.03	52.92	50.03	53.28	49.78
		Fraud Detection	CRA-CCF (CoT, Mcc)	0.83✓	-0.32	**2.73**	3.12	1.20	3.94	6.16
			CRA-CCFraud (CoT, Acc)	**34.03**✓	14.78	17.3	33.03	45.33	32.94	49.57
		Credit Scoring	Flare-German (CoT, Acc)	**64.00**✓	33.50	15.00	12.00	49.50	32.50	17.00
			Flare-Astralian (CoT, Acc)	44.60	**66.91**	11.51	12.95	46.76	56.12	51.80
			CRA-LendingClub (CoT, Acc)	**68.49**✓	52.69	25.38	23.40	48.87	21.03	65.03
		Distress Ident.	CRA-Polish (CoT, Mcc)	**15.30**✓	12.37	15.07	13.78	69.14	11.18	17.38
			CRA-Taiwan (CoT, Acc)	**40.81**✓	12.01	35.97	52.58	69.96	57.88	8.57
		Claim Analysis	CRA-ProroSeguro (CoT, Acc)	35.14	**96.98**	44.33	56.20	25.86	32.58	96.60
			CRA-TravelInsurance (CoT,Acc)	41.52✓	6.39	80.31	17.28	**94.48**	73.64	54.03
		Tabular QA	*Flare-TATQA (CoT, Acc)	**66.61**✓	63.43	63.70	64.21	57.70	66.40	74.90
		Open QA	*Finance Bench (CoT, Acc)	**54.00**✓	52.70	38.00	56.67	40.70	55.30	51.30
IF/Chat	General	Precise IF	MT-bench (1,2 turn avg)	7.36	7.88	**7.92**	5.80	8.38	7.84	9.10
Reasoning	Math	Math Reasoning	MathQA (CoT, Acc)	**55.08**✓	51.16	49.35	41.51	39.40	52.46	70.82
	General	Social Reasoning	Social-IQA (CoT, Acc)	**75.23**✓	68.83	70.73	77.28	72.82	62.95	78.92
		Common Sense	Open-book-qa (CoT, Acc)	**82.60**✓	77.00	82.20	87.00	80.20	76.40	94.60
			Hellaswag (CoT, Acc)	**81.90**✓	73.34	69.10	69.69	67.89	61.74	81.76
			Winogrande (CoT, Acc)	**70.32**✓	62.51	66.69	74.27	72.22	65.82	85.71
			PIQA (CoT, Acc)	**85.85**✓	79.82	81.45	86.72	82.05	77.91	94.34
	Finance	Exam	CFA-Easy (CoT, Acc)	**66.28**✓	60.56	60.47	36.05	61.24	65.89	83.14
			CFA-Challnge (CoT, Acc)	**55.56**✓	34.44	35.56	25.56	48.89	43.33	74.44

Table 3: Results on the **novel** tasks. The notations are the same as in Table 2. 'Mcc' refers to Matthews correlation coefficient, usually used in highly imbalanced data (Xie et al., 2024a).

FinDAP

Ablations on Preference Alignment

improved on 3 out 5

Task	Benchmark	Llama-Fin	Llama-Fin (w/o PA)
Sentiment Ana.	FPB	91.13	**92.99**
Sentiment Ana.	FiQA SA	**95.32**	94.47
Monetary Policy	FOMC	**64.31**	63.10
Named Entity	NER	**76.69**	74.33
Abs. Summ.	EDTSUM	53.78	**54.21**

Table 4: Ablation on PA on **similar (unseen)** evaluation set.

PA always improve

Mixed: some tasks are inherently 'easy' and reasoning capabilities might not be beneficial (important future work)

PA always improve

Capability	Domain	Task	Benchmark	Llama-Fin 8B	Llama-Fin (w/o PA)
Concept	General	Knowledge Recall	MMLU	**47.42**	47.22
			AI2-ARC	**89.43**	88.95
			Nq-open	**19.20**	16.20
	Finance	Knowledge Recall	MMLU-Finance	**64.20**	63.93
Task	Finance	Extract Summ.	Flare-ECTSUM	34.10	**34.41**
		ESG Issue	MLESG	40.67	**42.00**
		Rumor Detection	MA	84.00	**84.60**
		Stock Movement	SM-Bigdata	**54.14**	52.04
			SM-ACL	**51.99**	49.89
			SM-CIKM	**54.94**	44.88
		Fraud Detection	CRA-CCF	**0.83**	0.61
			CRA-CCFraud	**34.03**	32.32
		Credit Scoring	Flare-German	**64.00**	60.50
			Flare-Astralian	44.60	**51.80**
			CRA-LendingClub	**68.49**	65.96
		Distress Ident.	CRA-Polish	**15.30**	0.65
			CRA-Taiwan	40.81	**96.41**
		Claim Analysis	CRA-ProroSeguro	35.14	**86.57**
			CRA-TravelInsurance	41.52	**98.50**
		Tabular QA	*Flare-TATQA	**66.61**	66.43
		Open QA	*Finance Bench	**54.00**	52.00
IF/Chat	General	Precise IF	MT-bench	**7.36**	7.29
Reasoning	Math	Math Reasoning	MathQA	**55.08**	54.30
	General	Social Reasoning	Social-IQA	**75.23**	73.64
		Common Sense	Open-book-qa	**82.60**	79.20
			Hellaswag	**81.90**	78.92
			Winogrande	**70.32**	67.48
			PIQA	**85.85**	84.39
	Finance	Exam	CFA-Easy	**66.28**	62.31
			CFA-Challnge	**55.56**	35.56

Table 5: Abaltion on PA on **novel** evaluation set.

FinDAP

More details

- arXiv (including detailed ablation, hyper-parameters, prompts...):
  - **Demystifying Domain-adaptive Post-training for Financial LLMs**
  - https://arxiv.org/abs/2501.04961

- Github: https://github.com/SalesforceAIResearch/FinDAP

- HF (FinEval): https://huggingface.co/datasets/Salesforce/FinEval