LONG AND SHORT OF LLMS

MANISH SHRIVASTAVA

OUTLINE

- Language Models and Families
- AI in Finance
- LLMs in Finance (and some tasks)
- Bloomberg GPT: Case Study
- LLMs for Financial Documents Analysis
- LLM Adoption Concerns
- (TimeGPT)

What are Large Language Models (LLMs)?

Definition

Statistical models trained on extensive text corpora to predict probability distributions of word sequences (P(W) for sequence W=W1,W2...Wn).

Architecture

Transformer-based models with self-attention mechanisms (GPT decoder-only, BERT encoder-only, T5 encoder-decoder).

Capabilities

Enhanced language understanding and generation, enabling applications across industries with zero-shot learning and prompt-based adaptation.

Evolution of Language Models

From Statistical Models to Transformers

Language models have evolved from n-gram statistical models to RNN-based architectures like LSTM and GRU, culminating in the transformer architecture introduced in 2017. This evolution has been driven by advances in computational power, large-scale datasets, and novel neural network architectures that enable processing of longer sequences and more complex relationships.

N-gram Models

Early statistical models treating word sequences as Markov processes, limiting to (n-1) preceding words.

RNN-based Models

LSTM and GRU architectures capturing long-term dependencies in sequential text data.

Transformer Models

Self-attention mechanisms enabling parallel relationships between words with efficient large-scale training.

WHAT IS A TRANSFORMER?

- A Google Brain model.
  - Variable-length input
  - Fixed-length output (but typically extended to a variable-length output)
  - **No recurrence**
- Uses 3 kinds of attention
  - Encoder self-attention.
  - Decoder self-attention.
  - Encoder-decoder multi-head attention.

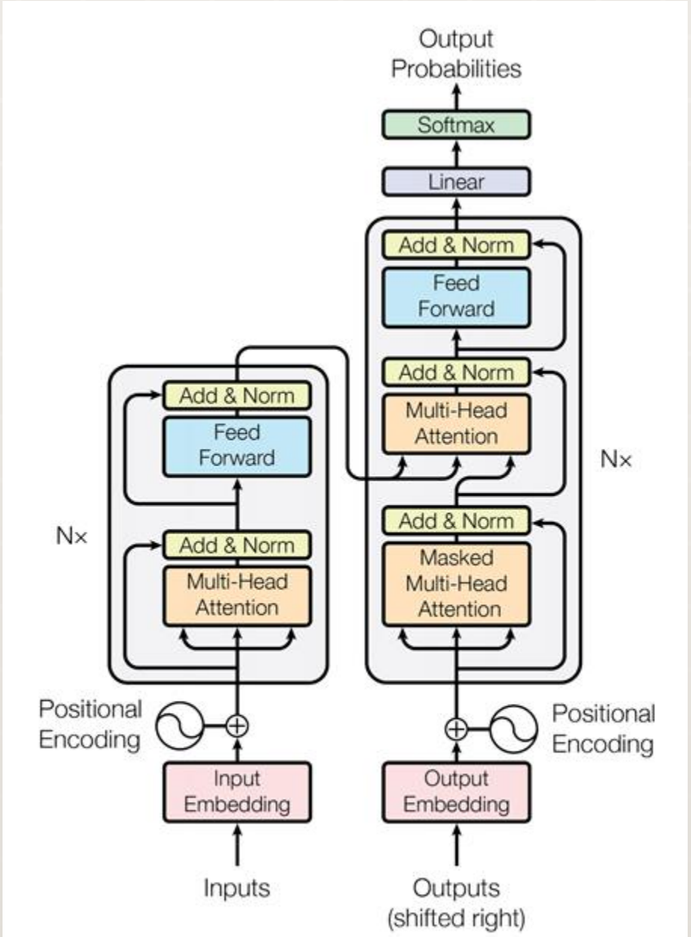


Figure 1: The Transformer - model architecture.

Transformer Architecture Overview

1 2 3

Step 1 - Input Processing

Text sequences converted to embeddings and positional encodings, transforming raw text into numerical representations that capture semantic meaning and word order relationships through learned vector spaces and sinusoidal positional encoding functions.

Step 2 - Self-Attention Mechanism

Words attend to all positions simultaneously, computing relationships and weights via scaled dot-product attention that dynamically determines the relevance of each word to every other word in the sequence, enabling context-aware representations.

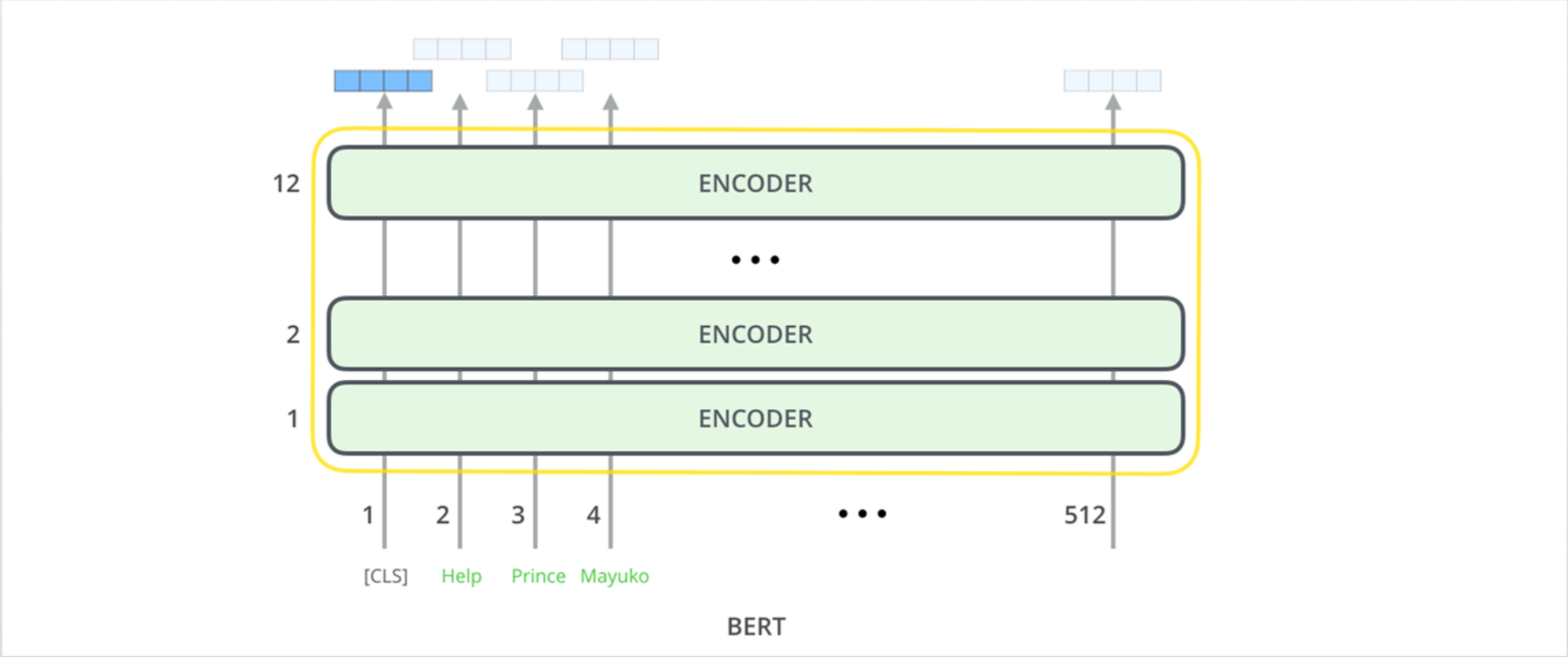
Step 3 - Output Generation

Attention outputs combined with feed-forward networks to produce predictions through layer normalization and residual connections, followed by linear transformations and softmax activation for final probability distribution over target vocabulary.

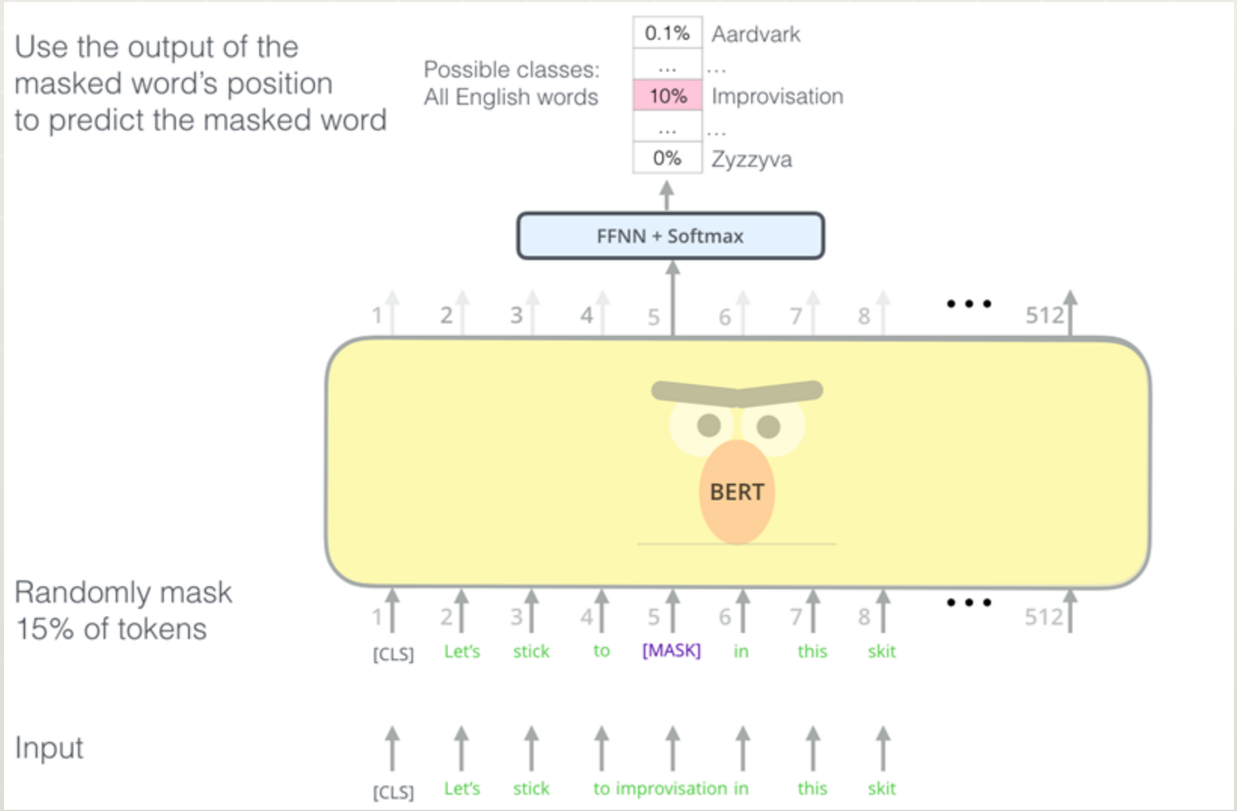
TWO MODEL FAMILIES: BERT AND GPT

- Children of Transformer
- BERT: Based on Transformer Encoder
  - Great at analysis
  - ‘Meh’ on Generation
- GPT: Based on Transformer Decoder
  - Good at analysis
  - Excellent at Generation

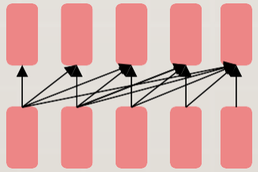
BERT'S ARCHITECTURE IS JUST A TRANSFORMER'S ENCODER STACK.



BERT IS TRAINED JUST LIKE A SKIP-GRAM MODEL.



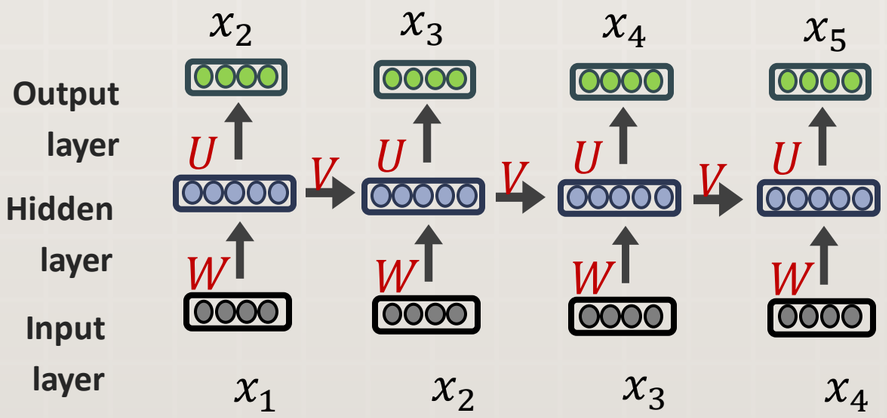
GPT



Decoders

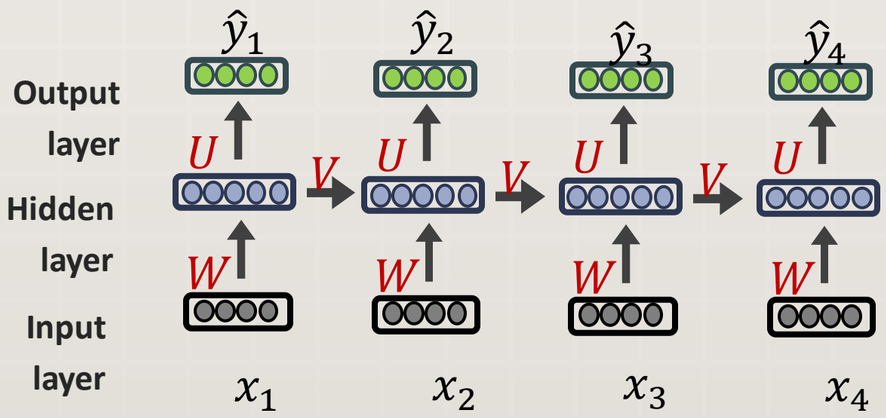
TERMINOLOGY: CAUSAL OR AUTO-REGRESSIVE MODEL

Language Modelling



Auto-regressive

1-to-1 tagging/classification



Non-Auto-regressive

GPT

Generative Pre-trained Transformer

GPT-2: A Big Language Model (2019)

Language Models are Unsupervised Multitask Learners

Alec Radford [*1] Jeffrey Wu [*1] Rewon Child [1] David Luan [1] Dario Amodei [**1] Ilya Sutskever [**1]

GPT: An Auto-Regressive LM (2018)

Improving Language Understanding by Generative Pre-Training

Alec Radford OpenAI alec@openai.com
Karthik Narasimhan OpenAI karthikn@openai.com
Tim Salimans OpenAI tim@openai.com
Ilya Sutskever OpenAI ilyasu@openai.com

GPT-2

- GPT-2 uses only Transformer Decoders (no Encoders) to generate new sequences from scratch or from a starting sequence

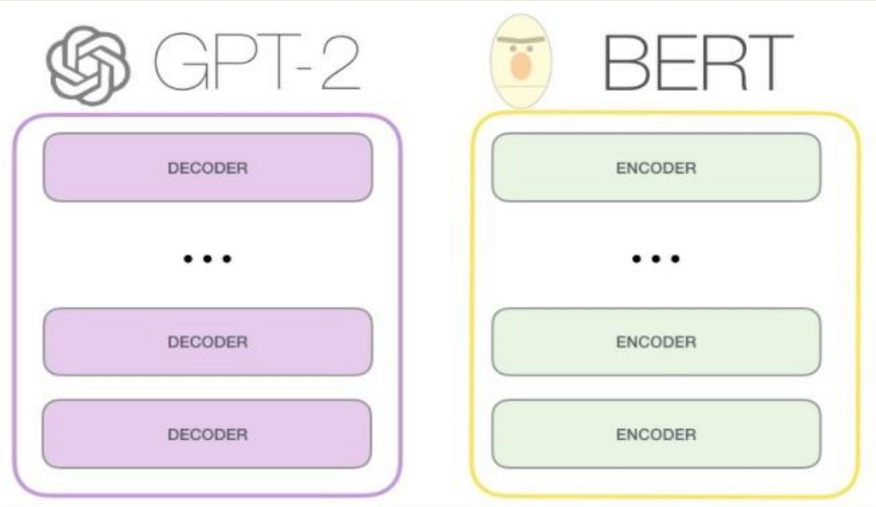


Image by http://jalammar.github.io/illustrated-gpt2/

GPT0-2: NEXT WORD PREDICTION

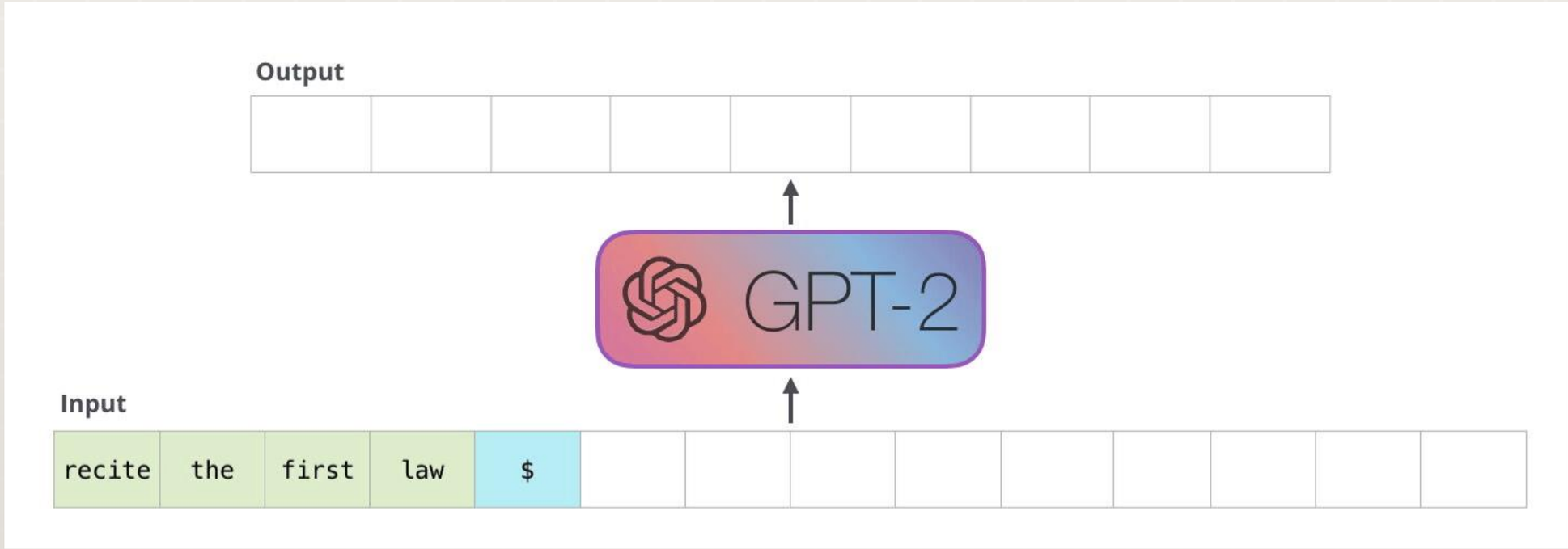


Image by http://jalammar.github.io/illustrated-gpt2/

GPT-2

- As it processes each subword, it masks the "future" words and conditions on and attends to the previous words

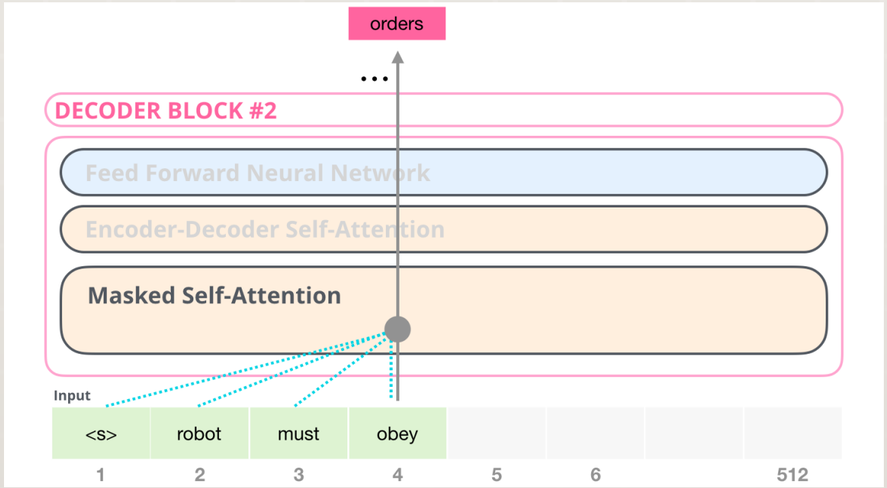


Image by http://jalammar.github.io/illustrated-gpt2/

GPT-2

- As it processes each subword, it masks the "future" words and conditions on and attends to the previous words

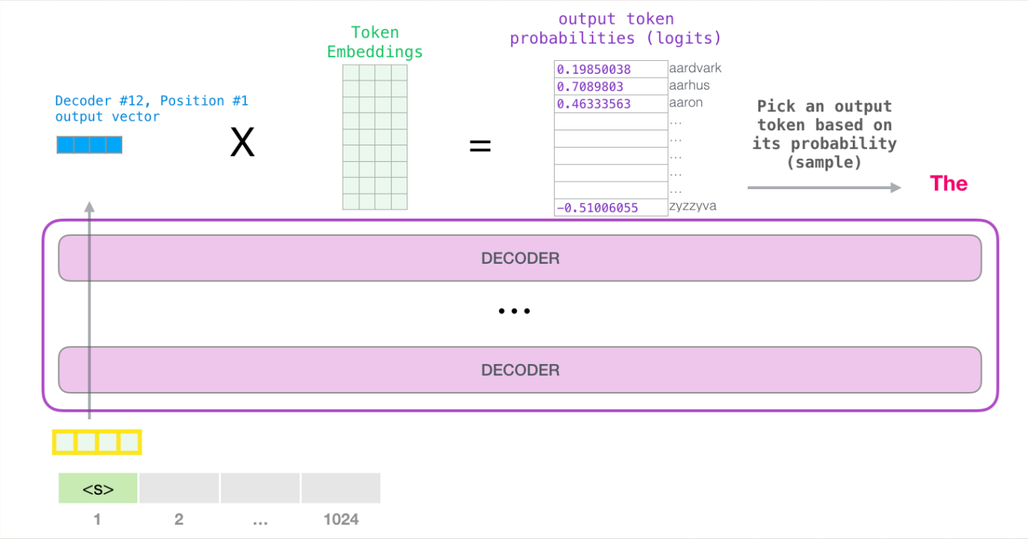
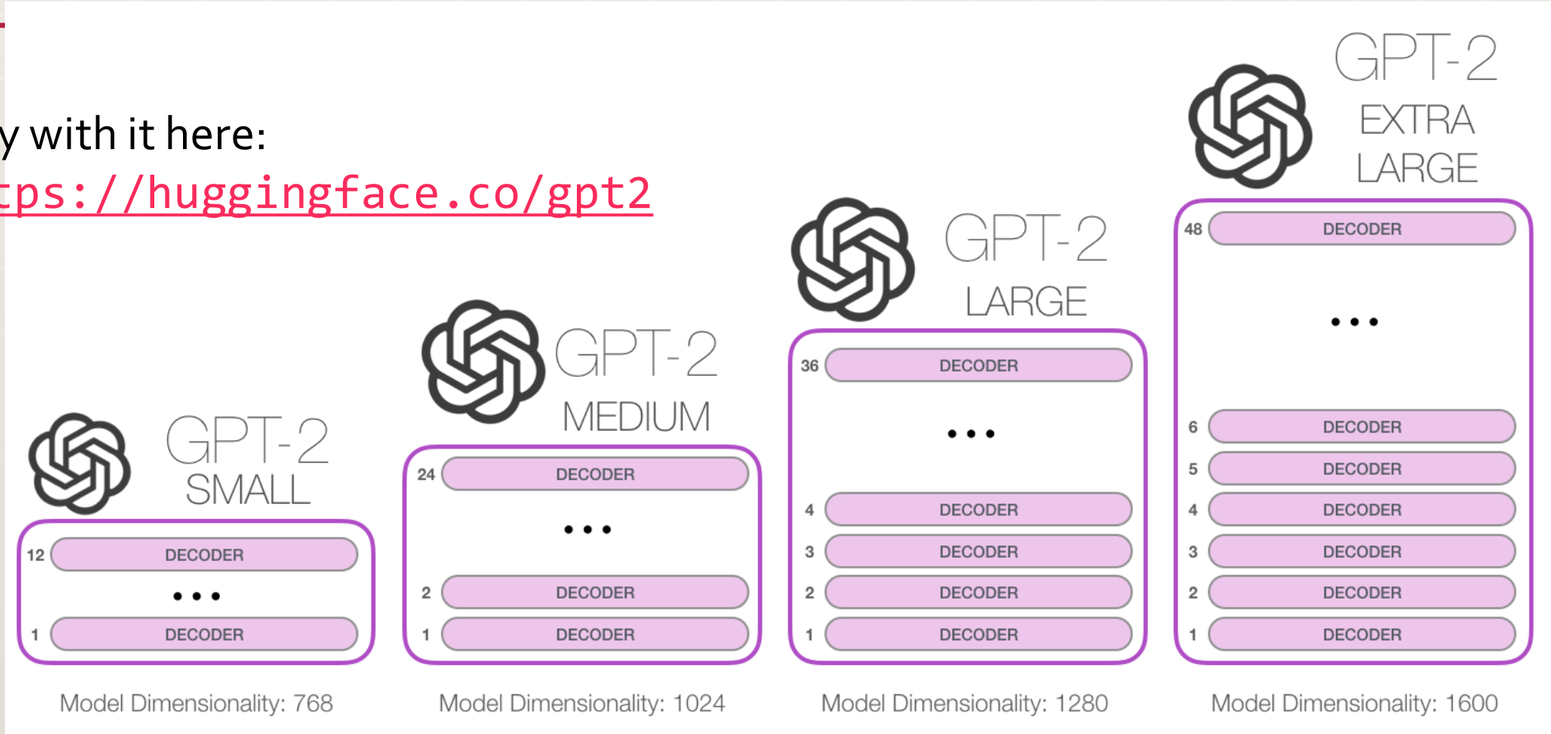


Image by http://jalammar.github.io/illustrated-gpt2/

GPT2: MODEL SIZES

Play with it here:
https://huggingface.co/gpt2



117M parameters 345M 762M 1542M

[Image by http://jalammar.github.io/illustrated-gpt2/]

AI Applications in Finance: Current Landscape

6

Primary Application Areas

100%

Industry Adoption Rate

5+

ML Techniques Used

2017

Transformer Inception Year

Trading & Portfolio Management

> Machine learning and deep learning algorithms optimize trading strategies and portfolio allocation through price forecasting and automatic trading systems.

- Statistical models (SVM, XGBoost, decision trees)
- Deep neural networks (RNN, LSTM, CNN, transformers)
- Reinforcement learning for automatic optimization

Financial Risk Management

> Advanced analytics identify and mitigate financial risks through sophisticated pattern recognition and predictive modeling.

- Fraud detection and anomaly identification
- Credit scoring and creditworthiness assessment
- Bankruptcy prediction and early warning systems

Text Mining & Advisory Services

> NLP and deep learning extract insights from unstructured data and power intelligent customer service solutions.

- Market sentiment analysis from financial news
- Information extraction from reports and disclosures
- AI-powered chatbots and investment advisory systems

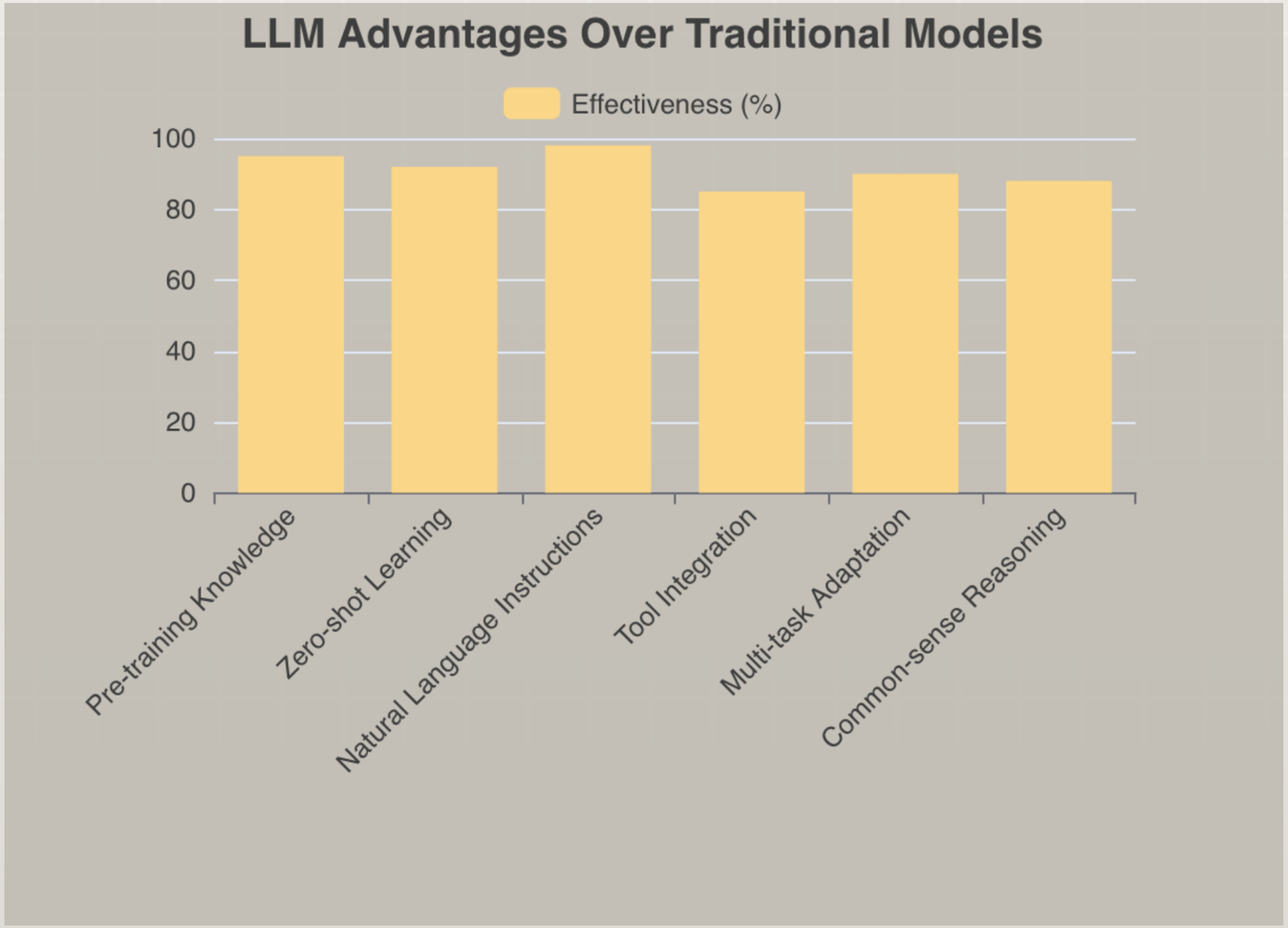
Advantages of LLMs in Financial Applications

Key Strengths

LLMs offer significant advantages over traditional machine learning models in finance in:

- **handling unstructured data and requiring minimal labeled training data.**
- **their ability to understand natural language instructions and perform zero-shot learning**

> **With pre-training knowledge and reasoning capabilities, LLMs can break down complex financial tasks into actionable steps automatically.**



LLMS IN FINANCE

FinBERT

- specialized large language model adapted for the finance domain
- Summarization, Sentiment Analysis etc

ChatGPT

- Effective for Research, Summarization, Generation
- Not specific to Finance

BloombergGPT

- 50 B parameter model
- Domain Specific
- Trained on General and Finance domain

FinGPT

- Better Than BloombergGPT
- 13B model (can be locally deployed)
- Public

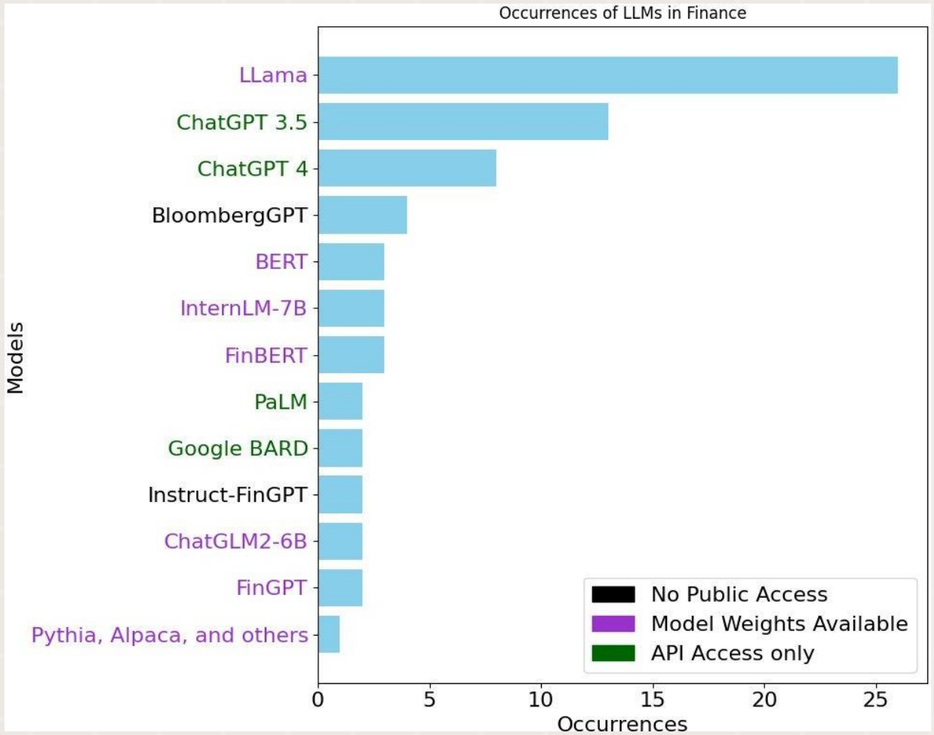
DeepSeek, Qwen, Llama etc...

- Public general purpose models
- Often better than Domain specific models

TimeGPT

- Trained on publicly available Time-Series data
- Not a text model, Time-series forecast only

LLAMA AND CHATGPT



Why is LLama at the forefront?

- Llama leads due to its adaptability and well-established ecosystem for fine-tuning, facilitating easier integration into financial applications.

Why is ChatGPT 3.5 more prevalent than 4?

- ChatGPT 3.5 enjoys greater usage over version 4 due to broader availability and cost-effectiveness, ensuring a balance between performance and accessibility.

Why is BloombergGPT used less than LLama?

- BloombergGPT may see less utilization compared to LLaMA because of limited access in comparison to open-source models.

COMMON TASKS

- Sentiment Analysis
- Text Classification
- Named Entity Recognition
- Relation Extraction

- Almost all use cases for LLMs in Finance are around
  - Text analysis
  - Text + Tabular analysis
  - Multimodal Analysis (Text + Image)
    - ChatGPT, Gemini lead the race here

Sentiment Analysis

- Financial PhraseBank
  - Sampled from financial news and company press releases.
  - Tagged as positive, negative or neutral by 16 annotators with domain knowledge.
  - E.g., Operating profit rose to EUR 13.1 mn from EUR 8.7 mn in the corresponding period in 2007 representing 7.7 % of net sales

positive

- FiQA
  - Include aspect-based sentiments from news in the financial domain

- TweetFinSent
  - Made with Tweets that contain retail investors' mood to a specific stock.
  - Created by AI Research
  - "Biggest up moves $COIN Coinbase, $PATH UIPath, $RBLX Roblox, $DKNG Draftkings, but $PLTR Palantir is stable"

positive

neutral

Data	50% Agreement		100% Agreement	
Model	Accuracy	F1 score	Accuracy	F1 score
ChatGPT(0)	0.78	0.78	0.90	0.90
ChatGPT(5)	0.79	0.79	0.90	0.90
GPT-4(0)	0.83	0.83	0.96	0.96
GPT-4(5)	**0.86**	**0.86**	**0.97**	**0.97**
BloombergGPT(5)	/	0.51	/	/
GPT-NeoX(5)	/	0.45	/	/
OPT66B(5)	/	0.49	/	/
BLOOM176B(5)	/	0.50	/	/
FinBert	**0.86**	0.84	**0.97**	0.95
ChatGPT(FT)	0.89	0.89		

Model	Accuracy	Weighted F1
ChatGPT(0)	68.48	68.60
ChatGPT(5)	69.93	70.05
GPT-4(0)	69.08	69.17
GPT-4(5)	71.95	**72.12**
ChatGPT((0_no_emoji))	64.40	64.43
ChatGPT((5_no_emoji))	67.37	67.61
GPT-4((0_no_emoji))	67.26	67.45
GPT-4((5_no_emoji))	70.58	70.44
RoBERTa-Twitter	**72.30**	71.96
ChatGPT(FT)	75.3	75.3

Table 4: Results on the TweetFinSent dataset.

Text Classification

- **Task: Financial News Headline Classification**
- **Data**: news headlines about gold commodity from various financial news provider sites (Reuters, The Hindu, The Economic Times, Bloomberg etc.)
- **Six categories**: price up, price down, price stable, past price, future price, and asset comparison

price up

1. Dec. gold settles at $1,293.80/oz, up $8.80, or 0.7%.

price down

2. Gold prices slide $14.90, or 1.1%, to $1,289.80 an ounce

past price

3. Gold imports dip 8% to $31.72 bn in 2015-16.

Model	Weighted F1
ChatGPT(0)	71.78
ChatGPT(5)	74.84
GPT-4(0)	84.17
GPT-4(5)	86.00
BloombergGPT(5)	82.20
GPT-NeoX(5)	73.22
OPT66B(5)	79.41
BLOOM176B(5)	76.51
BERT	**95.36**

Table 5: Results on the headline classification task.

ChatGPT(FT) 92.64

Name Entity Recognition

Task: FIN3 - NER

- **Data**: sentences extracted from SEC financial agreements
- **Four NER labels**: PER, LOC, ORG and MISC

- "tags": [0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
- Traditional approaches:
  - Rule-based methods
  - Statistical-modeling methods

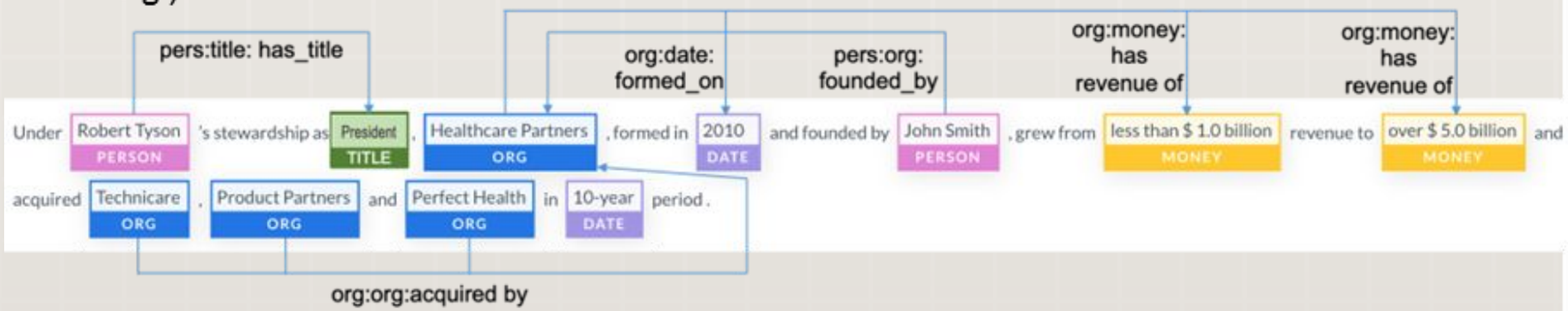
Model	Entity F1
ChatGPT(0)	29.21
ChatGPT(20)	51.52
GPT-4(0)	36.08
GPT-4(20)	56.71
BloombergGPT(20)	60.82
GPT-NeoX(20)	60.98
OPT66B(20)	57.49
BLOOM176B(20)	55.56
CRF(CoNLL)	17.20
CRF(FIN5)	**82.70**

Table 6: Results of few-shot performance on the NER dataset. CRF(CoNLL) refers to CRF model that is trained on general CoNLL data, CRF(FIN5) refers to CRF model that is trained on FIN5 data. Again, we choose the same shot as BloombergGPT for fair comparison.

Relation Extraction

Task: REFinD

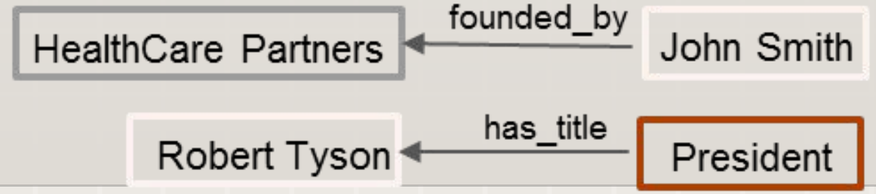
- **Data**: created from SEC 10-K/Q filings by AI Research
- **22 relation types**: org:org:subsidiary_of, org:org:agreement_with, etc.
- E.g.,



- Traditional approaches:
  - Rule-based methods
  - Unsupervised methods
  - Supervised methods

Model	Macro F1
ChatGPT (0)	20.97
ChatGPT (10)	29.53
GPT-4 (0)	42.29
GPT-4 (10)	46.87
Luke-base (fine-tune)	**56.30**

Table 7: Results on the REFinD dataset.



QUESTION ANSWERING

- **Tasks: FinQA and ConvFinQA**

- **Data**: Questions derived from earnings reports companies. Demands numerical reasoning and understanding of structured data and financial concepts. Emphasizes the ability to relate follow-up questions to previous conversation context.

Page 91 from the annual reports of GRMN (Garmin Ltd.)

The fair value for these options was estimated at the date of grant using a Black-Scholes option pricing model with the following weighted-average assumptions for 2006, 2005 and 2004:

	2006	2005	2004
Weighted average fair value of options granted	$20.01	$9.48	$7.28
Expected volatility	0.3534	0.3224	0.3577
Distribution yield	1.00%	0.98%	1.30%
Expected life of options in years	6.3	6.3	6.3
Risk-free interest rate	5%	4%	4%

... The total fair value of shares vested during 2006, 2005, and 2004 was $9,413, $8,249, and $6,418 respectively. The aggregate intrinsic values of options outstanding and exercisable at December 30, 2006 were $204.1 million and $100.2 million, respectively. *(... abbreviate 10 sentences ...)*

Question: Considering the weighted average fair value of options , what was the change of shares vested from 2005 to 2006?
Answer: - 400
Calculations:

$$\left(\frac{9413}{20.01}\right) - \left(\frac{8249}{9.48}\right) = -400$$

Program:

divide (9413, 20.01) | divide (8249, 9.48) | substract (#0, #1)

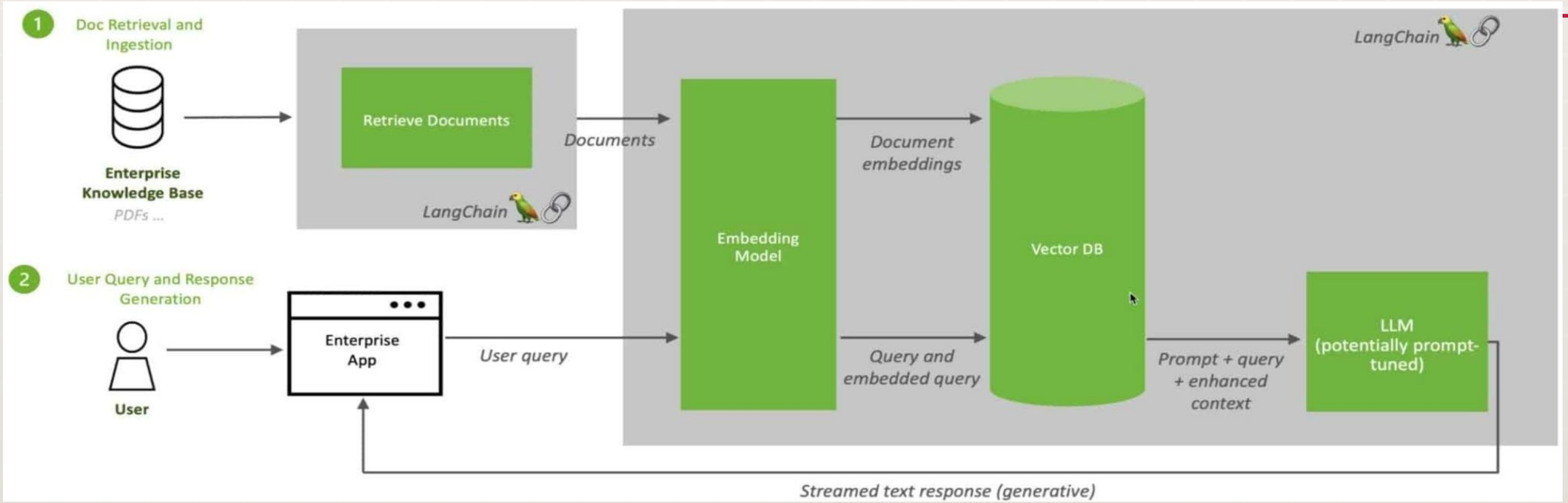
Model	FinQA	ConvFinQA
ChatGPT(0)	48.56	59.86
ChatGPT(3)	51.22	/
ChatGPT(CoT)	63.87	/
GPT-4(0)	68.79	**76.48**
GPT-4(3)	69.68	/
GPT-4(CoT)	**78.03**	/
BloombergGPT(0)	/	43.41
GPT-NeoX(0)	/	30.06
OPT66B(0)	/	27.88
BLOOM176B(0)	/	36.31
FinQANet(fine-tune)	68.90	61.24
Human Expert	91.16	89.44
General Crowd	50.68	46.90

Table 8: Model performance (accuracy) on the question answering tasks. FinQANet here refers to the best-performing FinQANet version based on RoBERTa-Large (Chen et al., 2022a). Few-shot and CoT learning cannot be executed on ConvFinQA due to the conservation nature of ConvFinQA.

- More advanced financial analysis QA:

Can GPT models be Financial Analysts? An Evaluation of ChatGPT and GPT-4 on mock CFA Exams, Callanan, E. et al, arXiv:2310.08678

RETRIEVAL AUGMENTED GENERATION (RAG)



Category of Questions	N	Questions (%)	GPT-3.5	GPT-4
Yes/No	143	42.43%	41.96%	46.85%
Single-Select from options	73	21.66%	38.36%	56.16%
Single-Select from options (number)	52	15.43%	17.31%	32.69%
Multiple-Select from options	43	12.76%	4.65%	25.58%
Number-extraction	26	7.72%	26.92%	19.23%

Accuracy of GPT3.5 and GPT4 on Cogtale dataset on different types of questions.

JZafaryab Rasool, Scott Barnett, Stefanus Kurniawan, Sherwin Balugo, Rajesh Vasa, Courtney Chesser, Alex Bahar-Fuchs (2023). Evaluating LLMs on Document-Based QA: Exact Answer Selection and Numerical Extraction using Cogtale dataset. arXiv:2311.07878 [cs.IR]

Training LLMs from Scratch: BloombergGPT Case Study

BloombergGPT Approach

Bloomberg trained a large language model from scratch combining public datasets with proprietary financial data. Despite using only a small percentage of finance-specific training data, BloombergGPT achieved significant performance improvements in finance benchmarks. This demonstrates that specialized financial pre-training, even on mixed datasets, substantially enhances domain-specific performance.

Hybrid Training Data

Combination of public corpora and proprietary Bloomberg financial datasets for comprehensive coverage.

Market Sentiment

Average score 62.51 vs. BLOOM176B 54.35 (+8.16 points improvement in classification).

Generative Capabilities

Significant outperformance in Question Answering, Named Entity Recognition, and summarization tasks.

Fine-Tuning Techniques for Finance Domain

Standard Fine-tuning

Training LLMs on raw financial datasets directly. Simple approach but may require more data.

Instruct Fine-tuning

Using task-specific datasets with examples and guidance. Better for understanding financial instructions and nuances.

LoRA & Quantization

Low-Rank Adaptation and parameter reduction techniques. ***Significantly lower computational requirements*** while maintaining performance.

Tool-Augmented Generation in Finance

Data Acquisition

LLMs can utilize Python packages and APIs to fetch real-time financial data, market prices, and economic indicators.

Portfolio Optimization

Tool-augmented LLMs access optimization libraries to construct efficient portfolios based on risk-return profiles.

Results Presentation

Automated generation of reports, visualizations, and actionable insights from analysis results.

Financial Documents Analysis

Efficiency gains

LLM pipelines convert thousands of pages into insights in minutes, dramatically accelerating financial analysis workflows.

Real-world applications

Tools like Talk to EDGAR, PitchBob, and BloombergGPT demonstrate practical implementations in financial services.

Performance challenges

Hallucinations remain, but specialized models consistently outperform generic ones in financial document analysis.

Human oversight necessity

AI serves as assistant, not final arbiter - human expertise remains essential for validation and decision-making.

Why Financial Documents Matter

Volume & Complexity

Average 10-K exceeds 150 pages with thousands of data points requiring labor-intensive manual review

Critical Decisions

Financial analysts, investors, regulators, and managers rely on accurate document analysis for investment decisions

Information Extraction Challenge

Dense prose, tables, technical terms, and boilerplate legal language make extraction difficult

The Financial Document Landscape

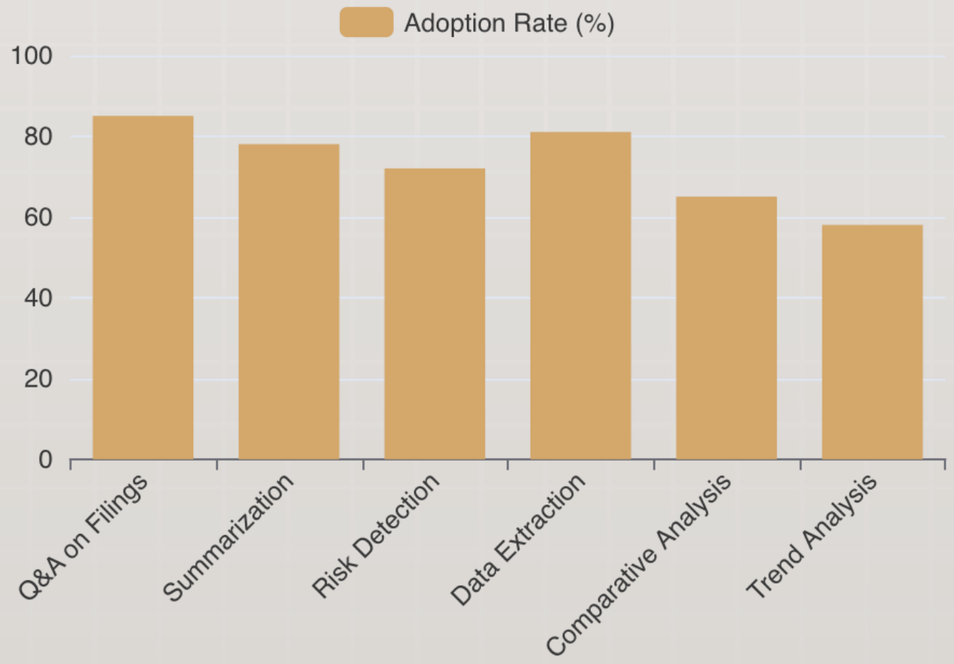
SEC Filings Overview

Forms like 10-K (annual), 10-Q (quarterly), and 8-K (current reports) are mandatory regulatory disclosures stored in the EDGAR database containing millions of filings spanning decades.

Corporate Decks

Investor presentations, pitch decks, earnings slides, and strategy documents are often 50-100+ slides combining text, images, charts and tables requiring intensive manual review.

Key Applications in Financial Analysis



Question Answering

Answer complex natural-language queries about filings in seconds, enabling analysts to extract facts without manual document review.

Risk Surveillance

Automatically detect, classify, and monitor risk factors, compliance issues, and emerging threats across multiple filings and time periods.

LLM Document Processing Pipeline

End-to-End Workflow

Documents are ingested through text extraction and OCR, segmented into logical chunks, converted to vector embeddings, and stored in searchable databases before LLM processing and retrieval-augmented generation.

Data Ingestion

Extract text from PDFs and PowerPoint slides using OCR tools, PDF parsers, and specialized vision models for handling multiple document formats.

Chunking & Indexing

Segment documents into logical sections, create semantic vector embeddings, and build searchable indexes for fast retrieval of relevant content.

LLM Processing

Apply retrieval-augmented generation using GPT-4, Claude, or fine-tuned finance models to answer queries, summarize, and extract structured insights.

SEC Filings Deep Dive: Key Document Types

Form 10-K

Annual comprehensive business overview with financial statements, MD&A, risk factors, legal proceedings, spanning 150+ pages

Form 10-Q

Quarterly report similar to 10-K but shorter, filed quarterly throughout the year

Form 8-K

Current report for unscheduled events like CEO changes, acquisitions, or earnings releases filed as needed

Other Forms

14A (proxy statements), SD (supply chain), S-1 (IPO registrations) serving specialized purposes

Real-World Application: SEC Filing Q&A Workflow

1

User Query

Analyst inputs natural language question about filing: "What were Company Y's revenue in Q2 2025 and reasons for change?"

2

Document Retrieval

System searches vector database and retrieves most relevant 10-Q sections via semantic similarity

3

Context Assembly

Retrieved snippets combined with query to form complete prompt for LLM

4

Answer Generation

GPT-4 or Claude generates structured answer citing exact text passages

5

Source Citation

System returns answer with hyperlinked quotes to original filing locations for verification

Question Answering: Query Types and Accuracy

Plain-Language Queries

What legal disputes did Company X mention? List top 5 risk factors. How much debt due in 2026? Success rate: 50-80% depending on query complexity

Fact Extraction

Find all merger references, extract numeric facts like net income. Structured generation (JSON/XML output) improves machine readability and validation

Table Reading

Questions requiring complex table interpretation often trip up LLMs. Solutions: specialized table-understanding modules or rerouting to Python tools for parsing

Numeric Accuracy

Challenges: misreading units, missing negative signs, hallucinating data. Solutions: context highlighting, retrieval validation, output verification against source text

Summarization Tasks: Methods and Results

Corporate Deck Summaries

LLMs generate five key takeaways, market analysis, or financial metrics from 50+ slide decks. Works best with clear, bullet-oriented original slides. GPT-4 achieves ~8/10 on coherence vs human summaries. Post-editing often required for refinement

SEC Filing Executive Summaries

Prompt: Summarize risk factors and MD&A sections for investor briefing. Challenges include number-heavy sections, legal language, and subtle qualifiers. Testing shows ~85% accuracy on identifying top risk categories

Real-World Example

Brightwave AI system summarized tech company quarterly reports identifying revenue growth drivers and cost pressures consistent with analyst consensus, though sometimes missing minor details

Performance Benchmarks & Accuracy

Benchmark/Task	Tool/Model	Accuracy/Result	
FinanceBench QA	GPT-4 Turbo with retrieval	19% accuracy (baseline)	-
FinanceBench QA	Custom RAG (LiveAI)	56% accuracy	-
10-K QA Benchmark	GPT-4	81.5%	-
PitchBob Analysis	AI vs VC Judges	High alignment on major categories	-
Processing Scale	MagicFinServ	70% cost reduction	

Future Directions

1

Larger Context Windows

models with 100K+ tokens enabling entire documents in-context

2

Company-Specific Fine-tuning

proprietary LLMs trained on organizational data for higher accuracy

3

Multimodal Integration

seamless analysis of text, tables, charts and video transcripts

4

Agentic Workflows

autonomous LLM agents performing multi-step analysis and generating reports

5

Regulatory Evolution

clearer standards for AI use in finance and investment advice

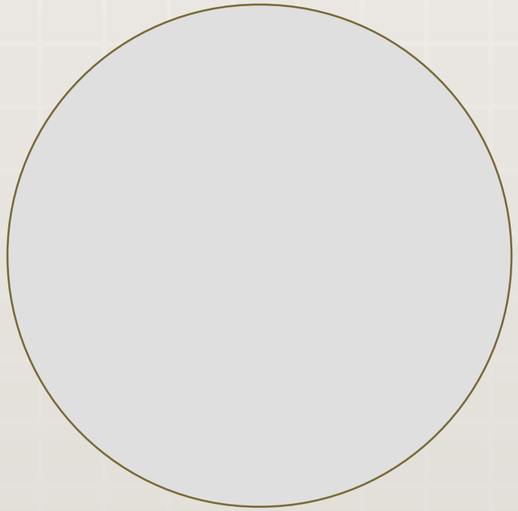
Multi-Agent Orchestration and Complex Workflows

Data Processing Agent

Retrieves filing chunks, performs OCR on images, extracts tables to structured form

Analysis Agent

Applies NER, sentiment analysis, compares metrics to historical data and forecasts



Validation Agent

Cross-checks facts against source documents, tests for hallucinations, escalates uncertain outputs for human review

Research Agent

Searches external data (news, earnings transcripts, analyst reports) to contextualize findings

Synthesis Agent

Generates summary reports, identifies implications, flags outliers or concerning trends

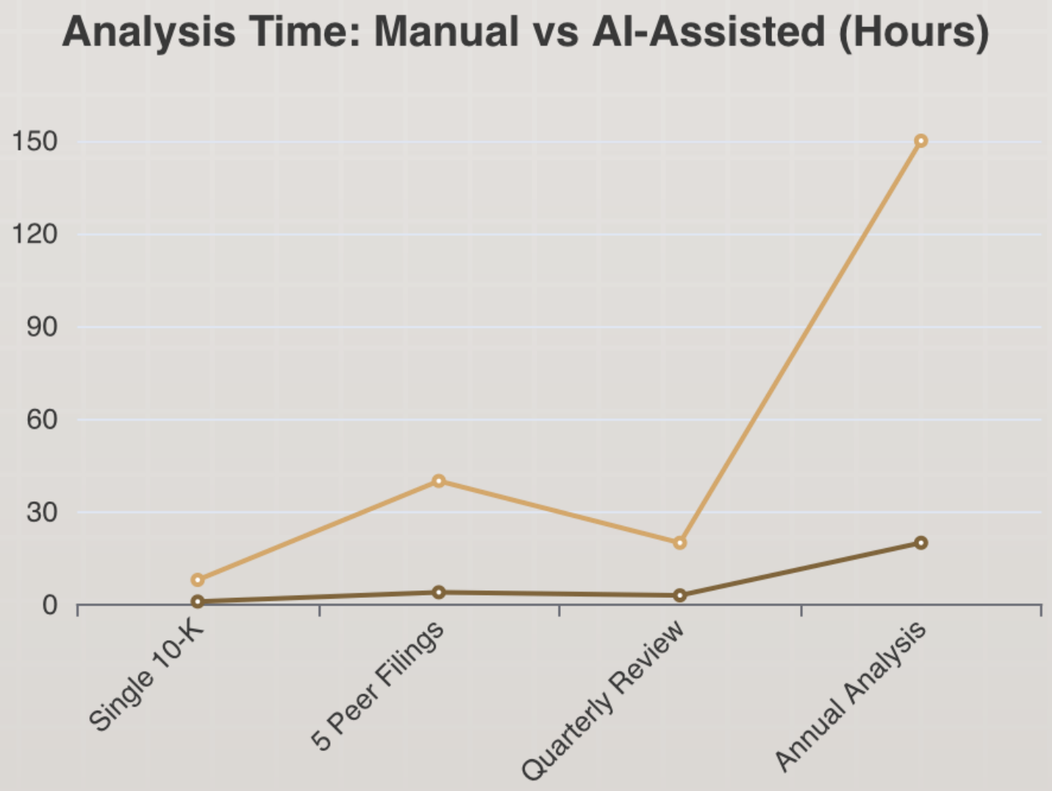
Speed Comparison: Manual vs AI Analysis

Dramatic Time Savings

Financial analysts traditionally spend hours reading, analyzing, and summarizing documents. AI-powered systems compress this workflow from days to minutes for routine analysis, freeing expert time for interpretation and strategic decision-making.

90%

Time Reduction Per Document



Efficiency Gains and Cost Metrics

Time Savings

MagicFinServ reports saving over 70% of time on filing reviews with AI assistance. Tasks taking expert teams hours/days completed in minutes. Talk to EDGAR generates instant answers vs clunky traditional EDGAR queries

Cost Reduction

Processing thousands of pages via AI costs fraction of hiring analysts. Per-document analysis cost drops from ~$500 expert review to ~$0.50 AI processing + light human oversight

Scale Enabling

AI makes analyzing peer groups or industry trends feasible. Comparing 50 company filings possible in hours instead of months

Analyst Productivity

Junior analysts write research notes 3-5x faster with AI summaries and pre-extracted data. More senior time available for interpretation and strategy rather than grunt work

Multi-Modal Challenges: Slides with Visual Content

Chart and Graph Extraction

Corporate decks contain visualizations showing market size, growth curves, competitive positioning. GPT-4o vision can interpret images but accuracy depends on clarity. Bar chart values sometimes misread, trend lines misinterpreted

Workflow Solutions

Convert charts to embedded tables before LLM processing. Use specialized computer vision models (Microsoft Document Intelligence, Adobe PDF API) for accurate data extraction. Maintain original images for human verification

Diagram Understanding

Organizational charts, product architecture diagrams, and flow charts require spatial reasoning beyond typical text analysis. Current LLMs struggle with complex multi-layer diagrams. Manual annotation or specialized models recommended for critical visuals

Domain Expertise and Prompt Engineering

Financial Jargon

Finance texts use domain-specific terms (GAAP, FFO, SG&A, EBITDA) that LLMs may misinterpret. Solutions: include glossaries in prompts, provide few-shot examples, use prompt templates specific to document type

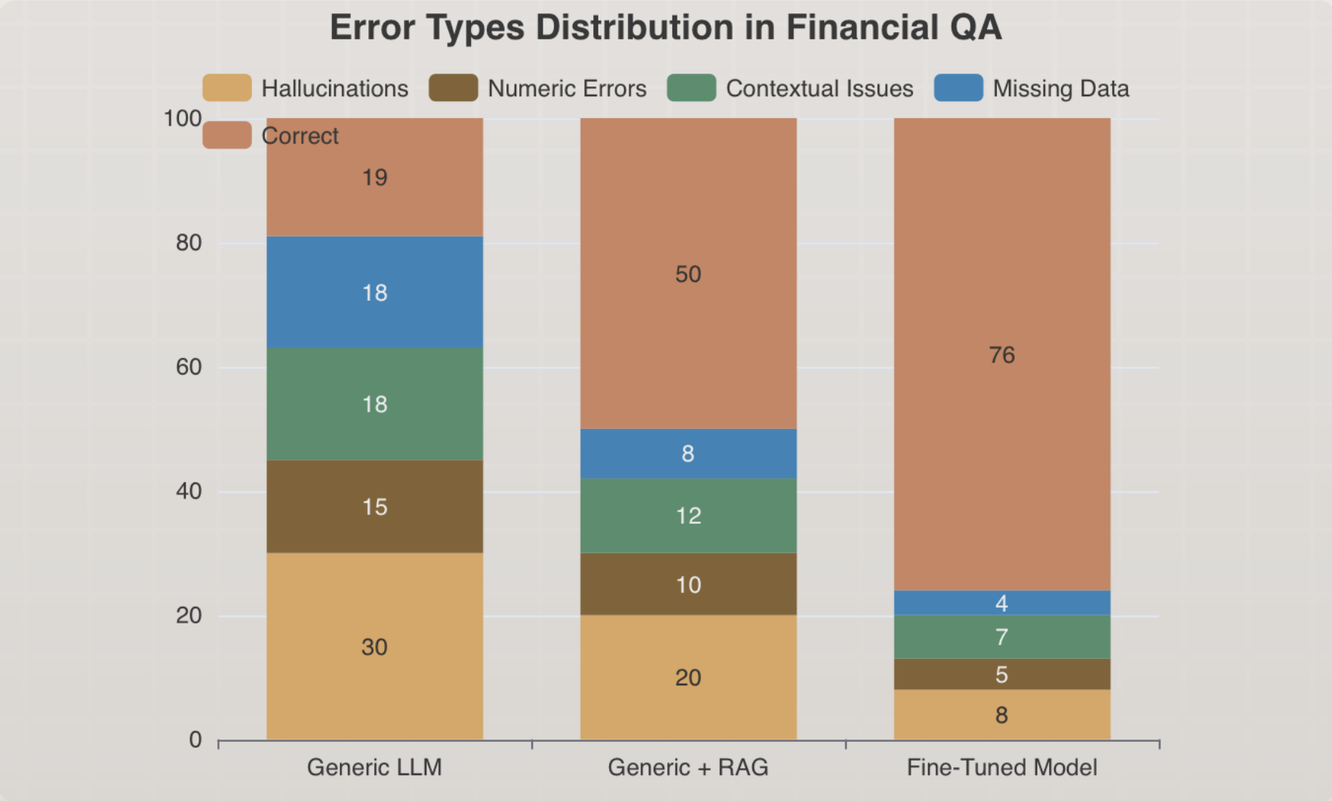
Context-Specific Instructions

Different document types require different analysis approaches. A 10-K MD&A prompt differs from pitch deck instructions. Pre-prompt templates encode best practices for consistent results

Few-Shot Learning

Providing 2-3 annotated examples of the task dramatically improves accuracy. Example: showing one correctly extracted risk factor teaches the model the expected format and detail level

Error Modes and Hallucination Rates



Mitigation Strategies

Reducing hallucinations requires multi-layered approaches combining retrieval optimization, model selection, and human oversight. Organizations implementing comprehensive validation frameworks achieve >90% accuracy on critical financial queries.

- **Require citations: Ground all answers in retrieved source text**
- **Highlight passages: Link output directly to supporting documents**
- **Confidence scores: LLMs output confidence levels for answers**
- **Fact-checking: Compare outputs against validated databases**
- **Human review: Critical decisions require expert validation**

Hallucinations: Error Modes

Fabricated Facts

ChatGPT invented a Wealthfront IPO slide that never existed in a pitch deck. Axios documented case where AI confidently generated false details about startup fundraising timelines

Numerical Errors

LLMs misread unit conversions (millions to billions), drop negative signs in losses, or report slightly different numbers due to token approximation rather than reading tables

Contextual Confusion

Misidentifying entities (accounting board tagged as company), confusing pronouns (attributing statements to wrong company), or missing negations (stating opposite of intended meaning)

Missing Context

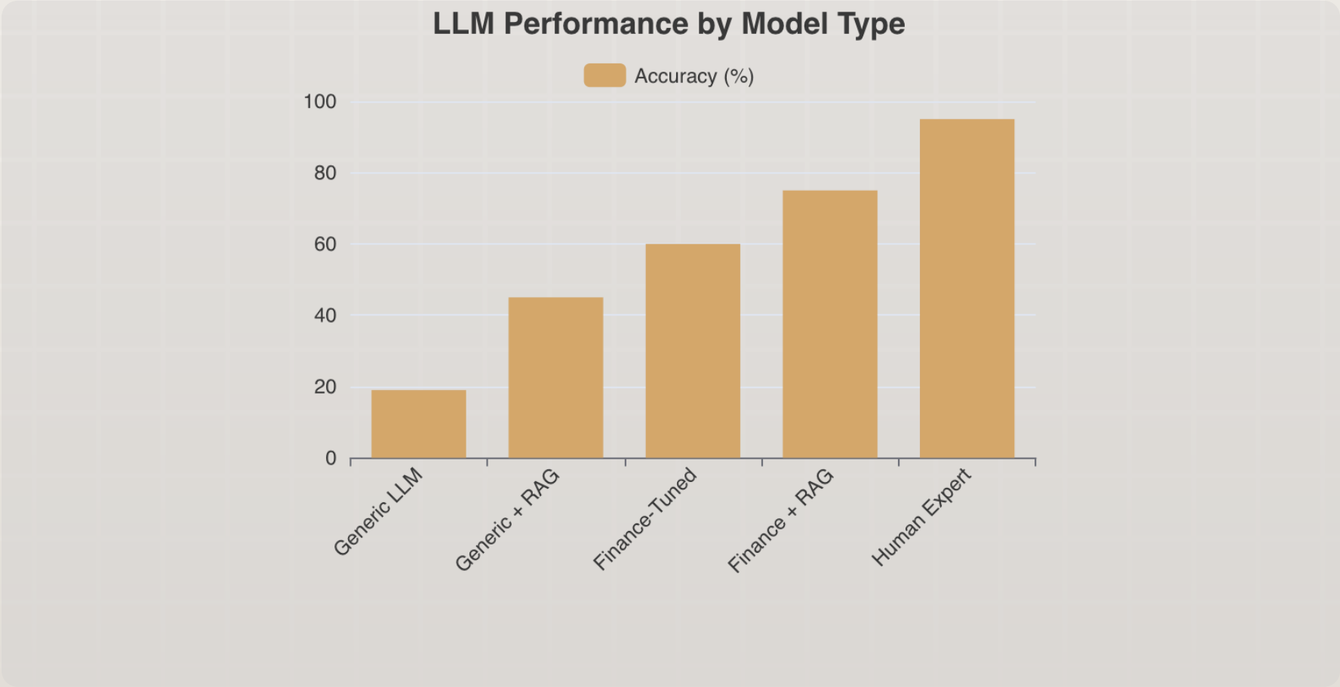
When retrieval fails to provide sufficient source text, LLMs may extrapolate plausibly but inaccurately. Insufficient context leads to confident wrong answers rather than 'I don't know' responses

Model Comparison Insights

Specialized financial LLMs significantly outperform generic models on domain-specific tasks. BloombergGPT trained on financial data achieves measurably higher accuracy on sentiment classification, named entity recognition, and question answering without degrading general capabilities.

> **FinBERT fine-tuned on earnings language demonstrates superior performance on MD&A analysis and forward-looking statement detection compared to standard BERT, confirming the value of domain-specific training.**

- **GPT-4 Turbo baseline: ~19% accuracy on FinanceBench QA**
- **Custom RAG systems: ~56% accuracy (3x improvement)**
- **Fine-tuned models: ~70-80% on specialized finance tasks**
- **Human analysts: ~95-100% on same tasks with source text**



Key Takeaway: LLMs as Financial Intelligence Accelerators

While AI excels at processing vast document volumes quickly, human expertise remains essential for validation and decision-making in regulated financial markets.

Conclusion: Key Findings Summary

Transformative Technology

LLM-based document analysis fundamentally changes how financial institutions process unstructured data. Scale and speed improvements enable analysis infeasible with traditional methods

Real-World Validation

Multiple case studies (Talk to EDGAR, PitchBob, BloombergGPT, MagicFinServ) demonstrate practical viability and significant efficiency gains in production environments

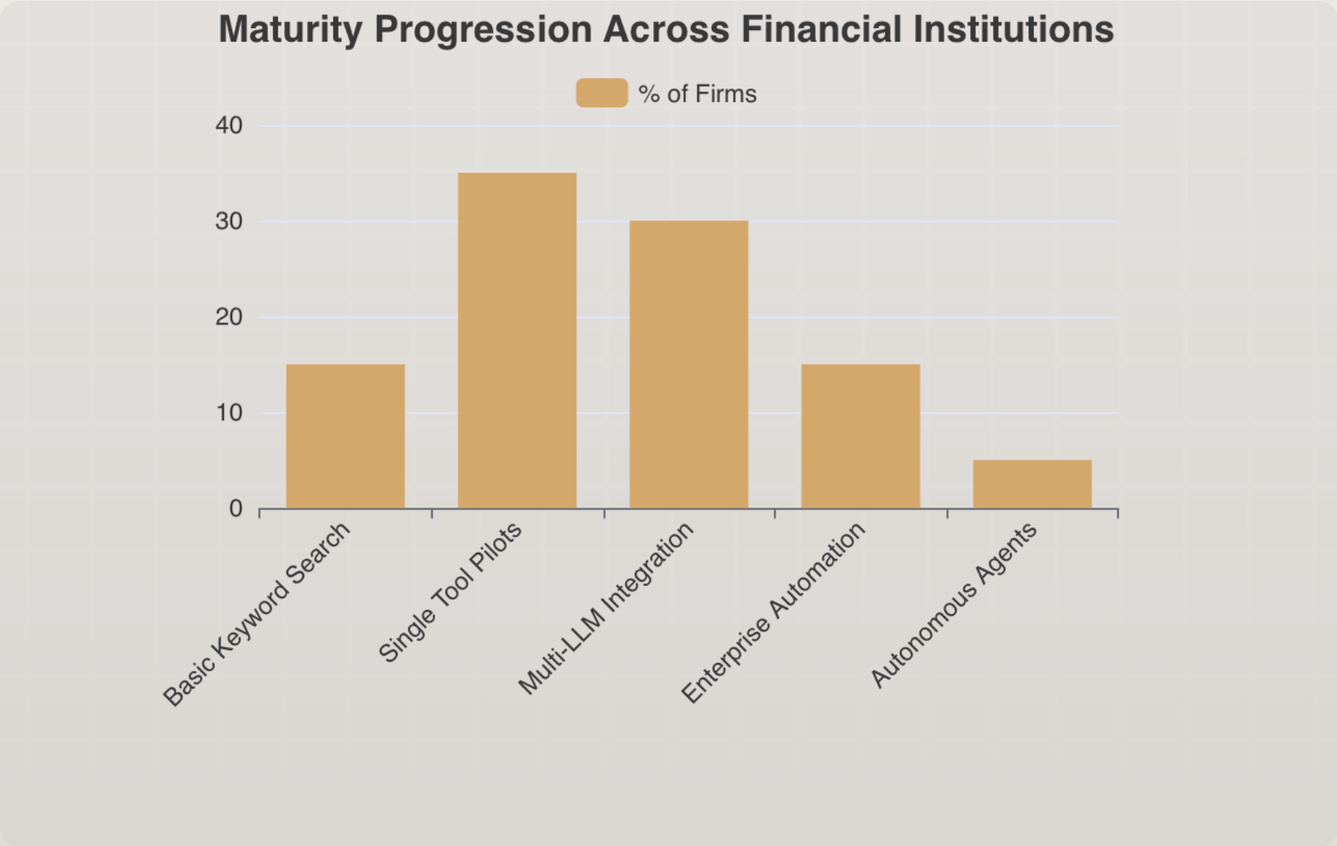
Limitations Are Clear

Hallucinations persist, multi-modal content remains challenging, and specialized finance training improves but doesn't eliminate errors. Human oversight remains essential

Strategic Imperative

Organizations must begin evaluating and piloting LLM solutions now. First-mover advantages exist for competitive intelligence, risk detection, and analyst productivity. Waiting risks competitive disadvantage

Adoption: From Early Adoption to Enterprise Scale



Capability Maturity

Most institutions still in early phases. Advanced firms have multi-tool integration and partial automation. Cutting-edge leaders developing proprietary agentic systems.

Market Trajectory

Rapid acceleration expected. As tools mature and competitive pressure increases, adoption will shift toward advanced automation and specialized models.

Key Decision Criteria: When to Use LLMs

Use LLM Insufficient annotated data + requires common-sense knowledge	S	W	**Use LLM** Unclear instructions + handles out-of-distribution data
Consider LLM Task needs multiple tool integration + complex reasoning required	T	O	**Use Traditional** Well-defined task + ample annotated data

Consider LLMs when pre-training knowledge, reasoning/emergent abilities, or orchestrating model collaboration are critical. Use conventional models if the task is well-defined with sufficient labeled data and minimal common-sense requirements.

Build vs Buy: Technology Options

Open Source Stack

Download Llama2/3, fine-tune on financial data, build RAG with LangChain, deploy on internal servers.

Advantage: full control and low per-query costs.
Disadvantage: requires ML expertise and infrastructure management

Commercial APIs

Use OpenAI GPT-4, Anthropic Claude, Google Gemini with API keys.

Advantage: simplicity, cutting-edge models, no infrastructure.

Disadvantage: recurring costs, data privacy concerns, vendor lock-in

Specialized Platforms

Talk to EDGAR, AlphaSense, BloombergGPT, PitchBob etc.

Advantage: pre-built domain knowledge, proven workflows.

Disadvantage: limited customization, high costs, must accept

Hybrid Approach

Use commercial LLM APIs for core intelligence, supplement with custom fine-tuning for proprietary data, integrate multiple tools for different tasks.

Optimal for many enterprises balancing cost, capability, and control

The Open vs Proprietary Model Debate

Open Models Advantages

Transparency enables bias auditing and control.
Reduced vendor lock-in.
Flexibility for customization on proprietary data.
Lower long-term costs at scale.
Community contributions and improvements

Proprietary Models Advantages

Cutting-edge capabilities (GPT-4, Claude 3, BloombergGPT).
Strong performance out-of-the-box. Vendor support and SLAs.
Specialized finance training from Bloomberg, data providers.
Reduced customization burden

Industry Evolution

Large institutions build proprietary finance LLMs internally.
Mid-market uses commercial APIs.
Small firms use specialized platforms.
Hybrid strategies combining multiple approaches will dominate

LLM Solution Deployment Options

Two Primary Approaches

Organizations can access LLM capabilities through third-party API providers or deploy open-source models locally. The choice depends on data privacy requirements, computational resources, performance expectations, and cost considerations. Third-party APIs offer simplicity but raise data confidentiality concerns, while self-hosted solutions provide greater control and privacy.

Proprietary APIs

OpenAI (GPT-3.5/GPT-4), Google (BARD), Microsoft. Features: chat, SQL generation, code completion. Cost-based on API calls.

Open-Source Models

LLaMA, BLOOM, Flan-T5, OpenLLaMA, Alpaca, Vicuna. Greater flexibility and privacy. Data remains under user control with accessible model weights.

Self-Hosting Requirements

For confidential data: NVIDIA V100 GPU (7B models), NVIDIA A100/A6000 (13B models). Robust machines required for local deployment.

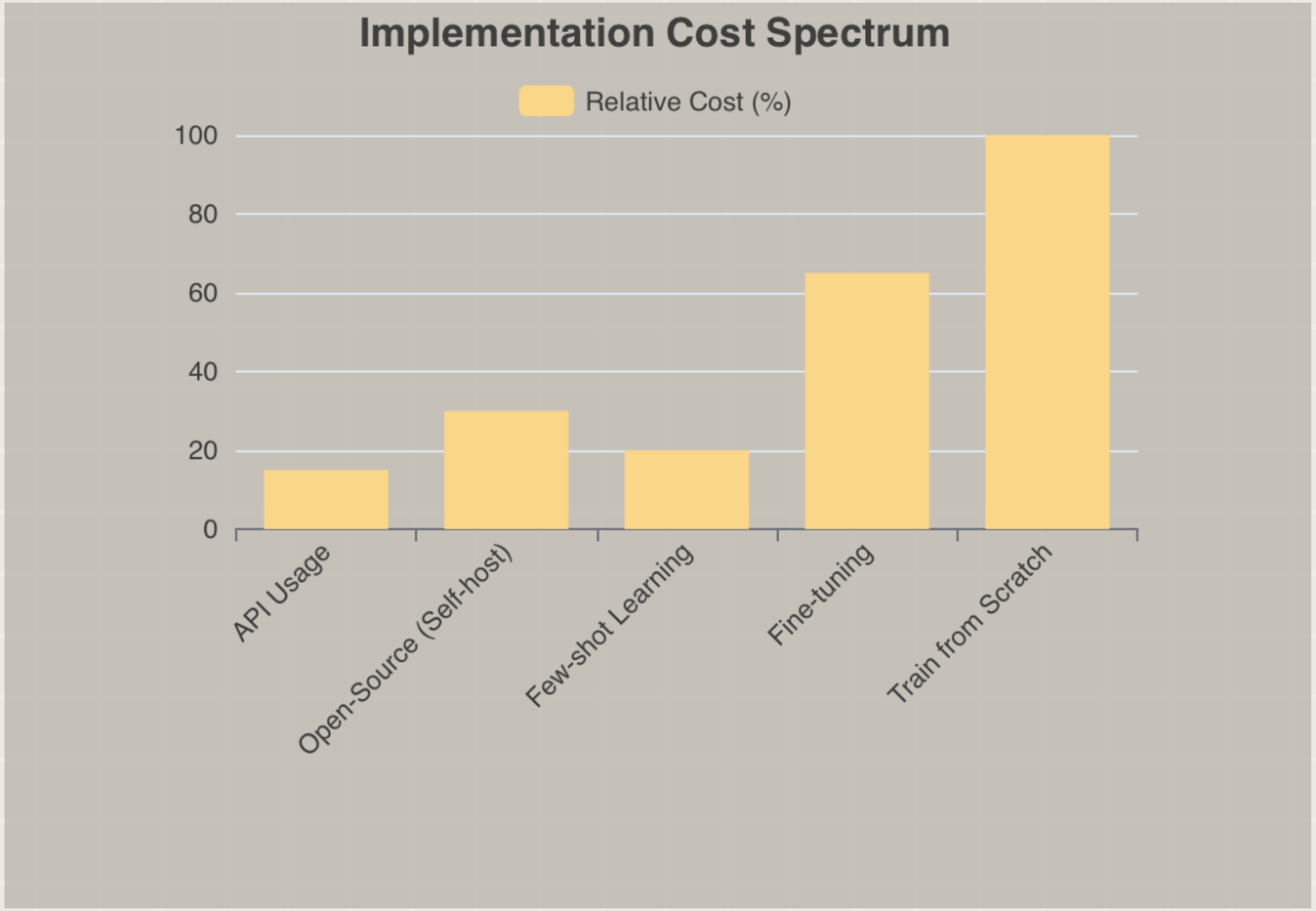
Cost-Benefit Analysis: Deployment Options

Implementation Costs

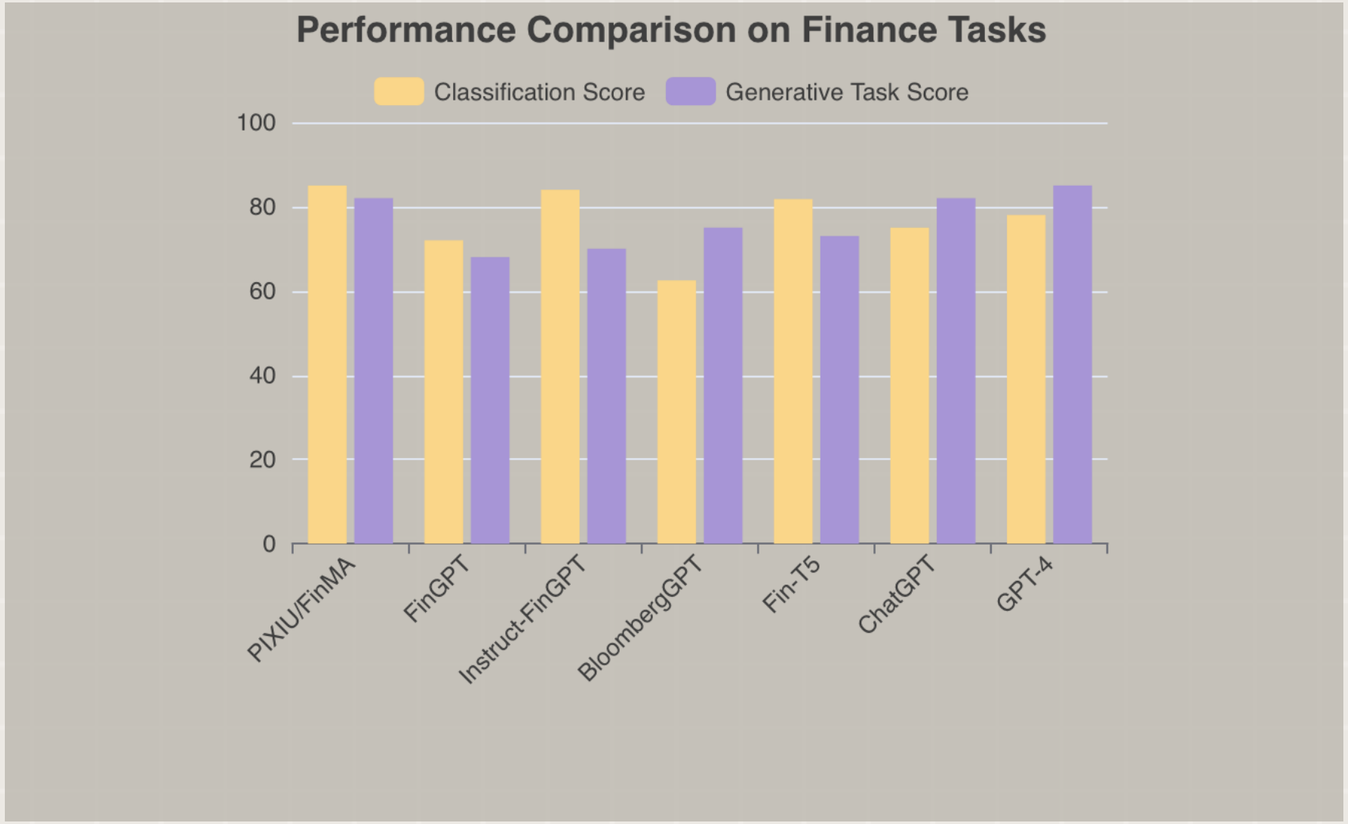
The financial investment in LLM deployment varies significantly based on the chosen approach. Third-party APIs offer low upfront costs but recurring expenses, while self-hosting requires infrastructure investment. Fine-tuning and training from scratch demand substantial computational and human resources. Careful cost-benefit analysis is essential for optimal resource allocation.

$$$

Costs increase significantly with complexity



Fine-Tuned Finance LLMs Performance

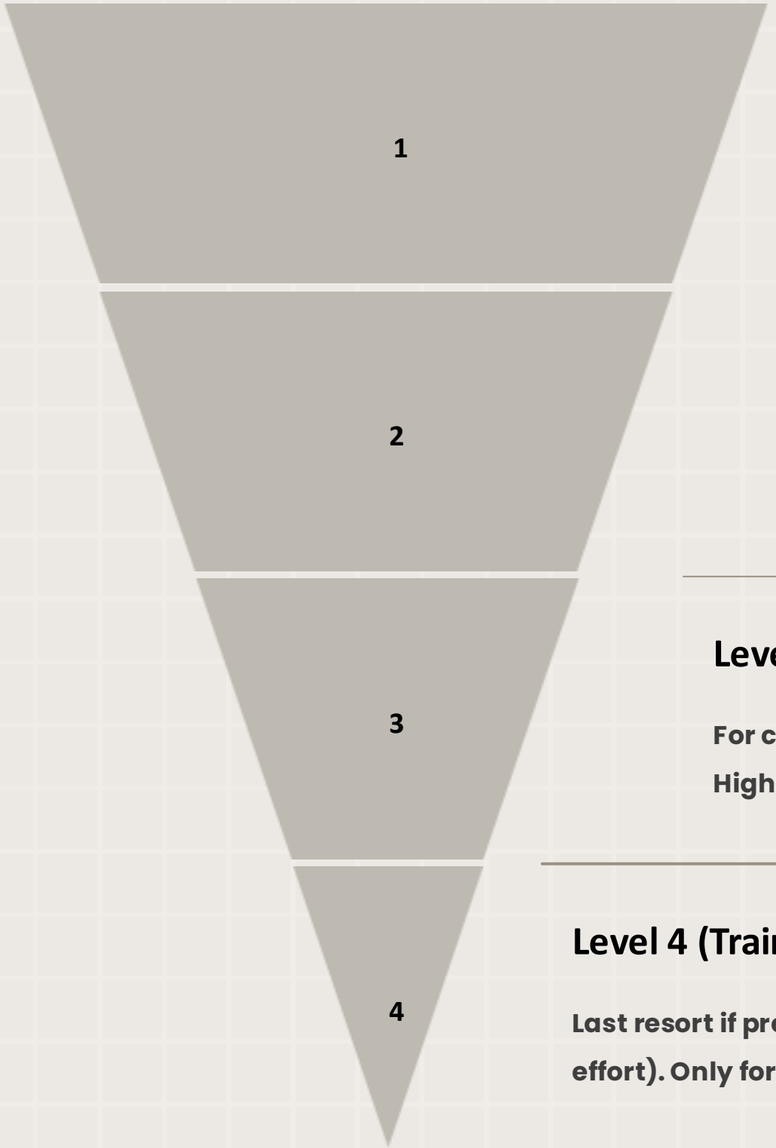


Key Insights

Fine-tuned finance-specific LLMs demonstrate superior performance in classification tasks compared to general-purpose models, achieving 81-85% accuracy. However, in generative tasks like question answering and summarization, general models like ChatGPT and GPT-4 maintain competitive advantages.

- **FinMA and Instruct-FinGPT lead in classification (82-85%)**
- **BloombergGPT: 62.51 avg (vs BLOOM176B: 54.35) - 8.16 point gain**
- **Fin-T5: 81.78 avg (vs T5: 79.56) - 2.22 point improvement**
- **Trade-off: excellent classification, moderate generation**

Four-Level Decision Framework for LLM Implementation



Level 1 (Zero-shot)

Use existing LLM services (GPT-3.5, GPT-4, BARD) or self-host open-source models (LLaMA, Alpaca) for lightweight experiments. Cost: Low (API calls or GPU resources).

Level 2 (Few-shot)

Provide 1-10 examples as context to improve performance. Similar cost to zero-shot with minimal additional overhead.

Level 3 (Tool-Augmented & Fine-tuning)

For complex tasks: integrate external tools/plugins or fine-tune with annotated data. Cost: Medium-High (tool development, computational resources, expertise).

Level 4 (Training from Scratch)

Last resort if previous levels fail. Cost: Extremely High (millions of dollars, trillions of tokens, months/years of effort). Only for mission-critical applications.

Implementation Best Practices

Assessment

Evaluate your financial task against the decision framework to determine if LLM is appropriate.

Selection

Choose between proprietary APIs or open-source based on data sensitivity and resource availability.

Experimentation

Start with zero-shot/few-shot approaches before investing in fine-tuning or custom solutions.

Validation

Rigorously test models on domain-specific financial benchmarks and real-world scenarios.

Deployment

Implement with monitoring, guardrails, and fallback mechanisms for production reliability.

Addressing LLM Limitations in Finance

Hallucination & Accuracy

Problem - LLMs may generate false information.Solution - Implement Retrieval-Augmented Generation (RAG) to ground responses in verified financial data sources.

Bias & Fairness

Problem - Manifestation of racial, gender, religious biases in outputs.Solution - Content censoring, output restrictions, and bias detection mechanisms during model training and inference.

Explainability

Problem - Black-box nature limits understanding of decision-making.Solution - Implement interpretability methods and maintain audit trails of critical financial decisions.

TimeGPT: The Challenge of Time Series Forecasting

Cross-Sector Importance

Time series data is essential across economics, energy, finance, and healthcare sectors

Methodological Divide

The forecasting community is divided on deep learning effectiveness compared to traditional approaches

Historical Dominance

Statistical and traditional ML methods have long dominated time series forecasting

Paradigm Shift

TimeGPT represents a fundamental transformation in forecasting methodologies

Background: Evolution of Forecasting Methods

Statistical Methods

Traditional approaches including ARIMA, ETS, and MSTL dominated time series forecasting for decades with their interpretable models and reliable performance

Machine Learning Era

The rise of XGBoost and LightGBM demonstrated superior performance on many forecasting tasks, combining feature engineering with powerful ensemble techniques

Deep Learning Debate

While RNNs, CNNs, and Transformers show promise, skepticism remains about their computational cost and lack of interpretability compared to traditional methods

Deep Learning in Time Series

1

Early Adapters

RNNs (DeepAR, ESRNN) pioneered deep learning applications in time series forecasting, establishing baseline architectures for sequential data processing.

2

CNN-based

Models like DPMN and TimesNet introduced convolutional approaches, capturing local patterns through hierarchical feature extraction.

3

Feed-forward

Architectures such as N-BEATS and NHITS demonstrated the effectiveness of purely feed-forward networks with interpretable basis expansions.

4

Transformer Era

The TFT, MQ-Transformer, and PatchTST family adapted attention mechanisms for temporal dependencies, while Informer/Autoformer addressed long-sequence efficiency.

5

Future Directions

Foundation models remain underexplored but show significant promise for unified time series representation learning across domains.

Foundation Models: Core Concepts

Foundation models generalize across domains and tasks. Transfer learning applies knowledge from source to target tasks. TimeGPT leverages scaling laws on dataset and model sizes. Key insight: Diversity enables better temporal pattern learning.

Generalization Across Domains

Foundation models transfer knowledge

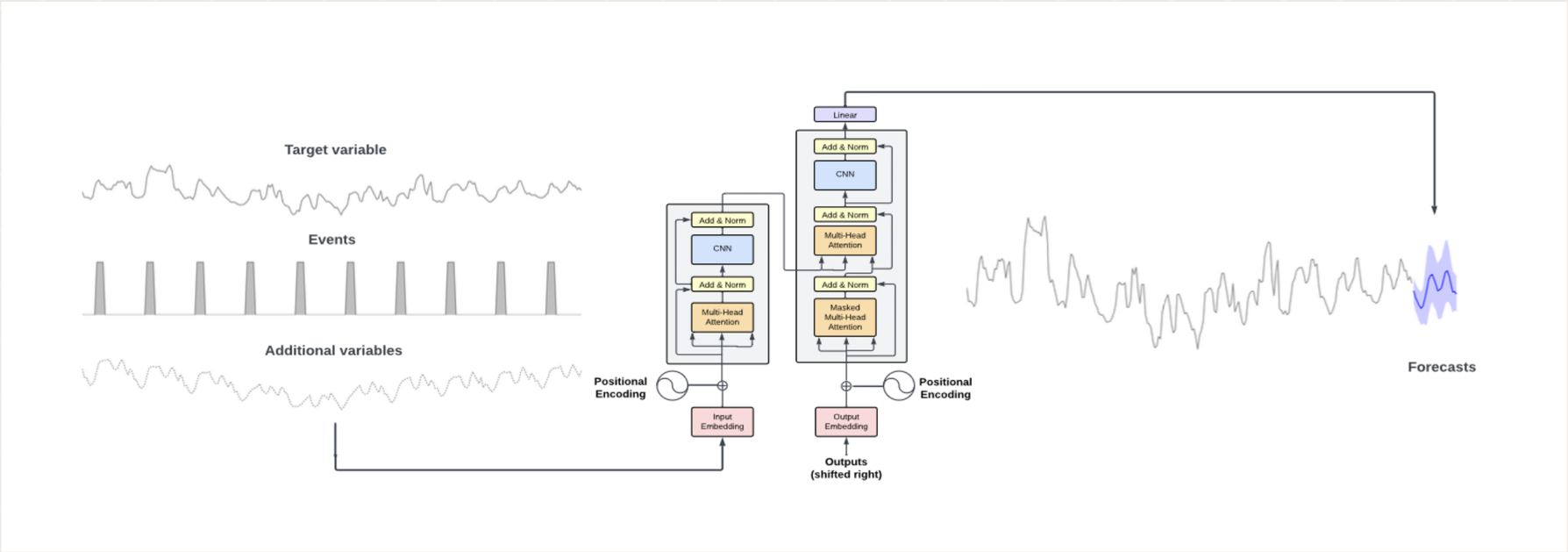
Scaling Laws

Larger datasets and models perform better

Diverse Patterns

Wide variety of temporal data improves robustness

TimeGPT Architecture: Technical Design



1

Input Layer

Local positional encoding enriches Time-Series input, handles varied frequencies with adaptive scaling mechanisms

2

Encoder

Multi-layer transformer with self-attention mechanisms, 12 parallel attention heads per layer

3

Decoder

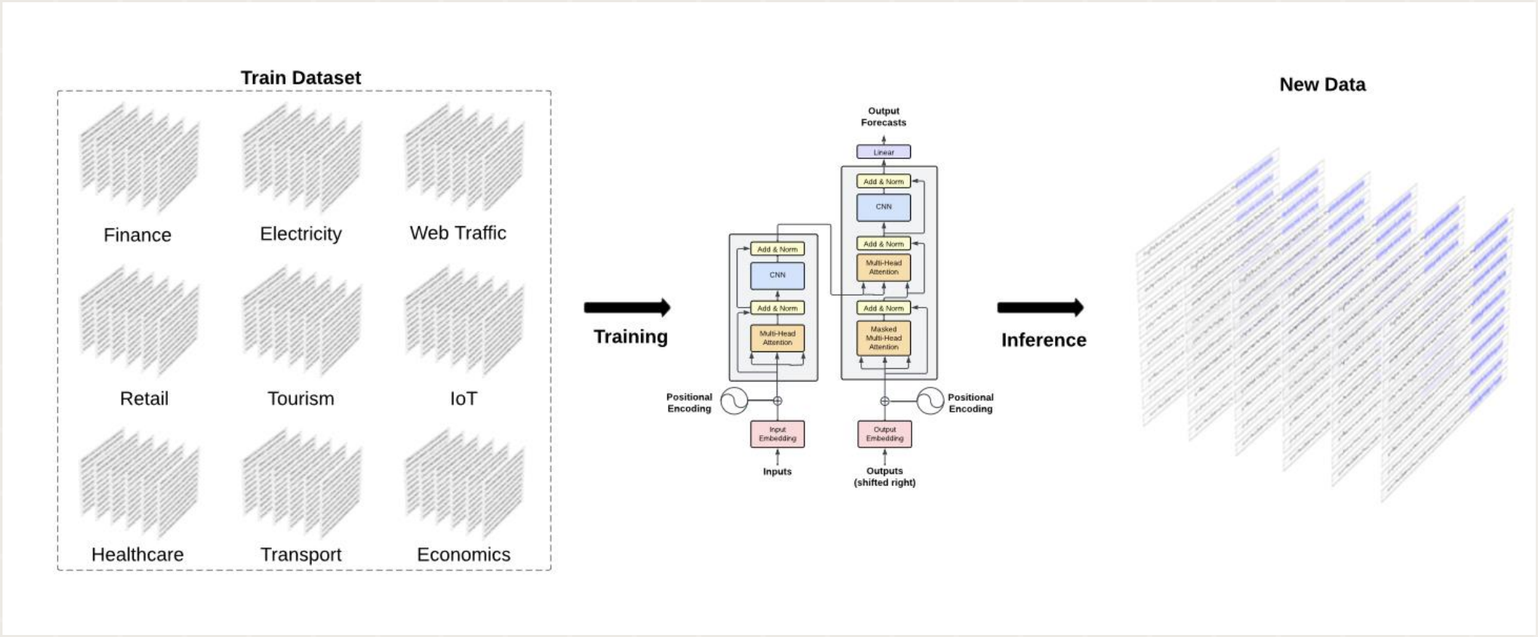
Residual connections and layer normalization, skip connections between every 3 transformer blocks

4

Output Layer

Linear layer maps to forecasting window dimension, dynamic output scaling based on prediction horizon

Training Dataset: Scale and Diversity



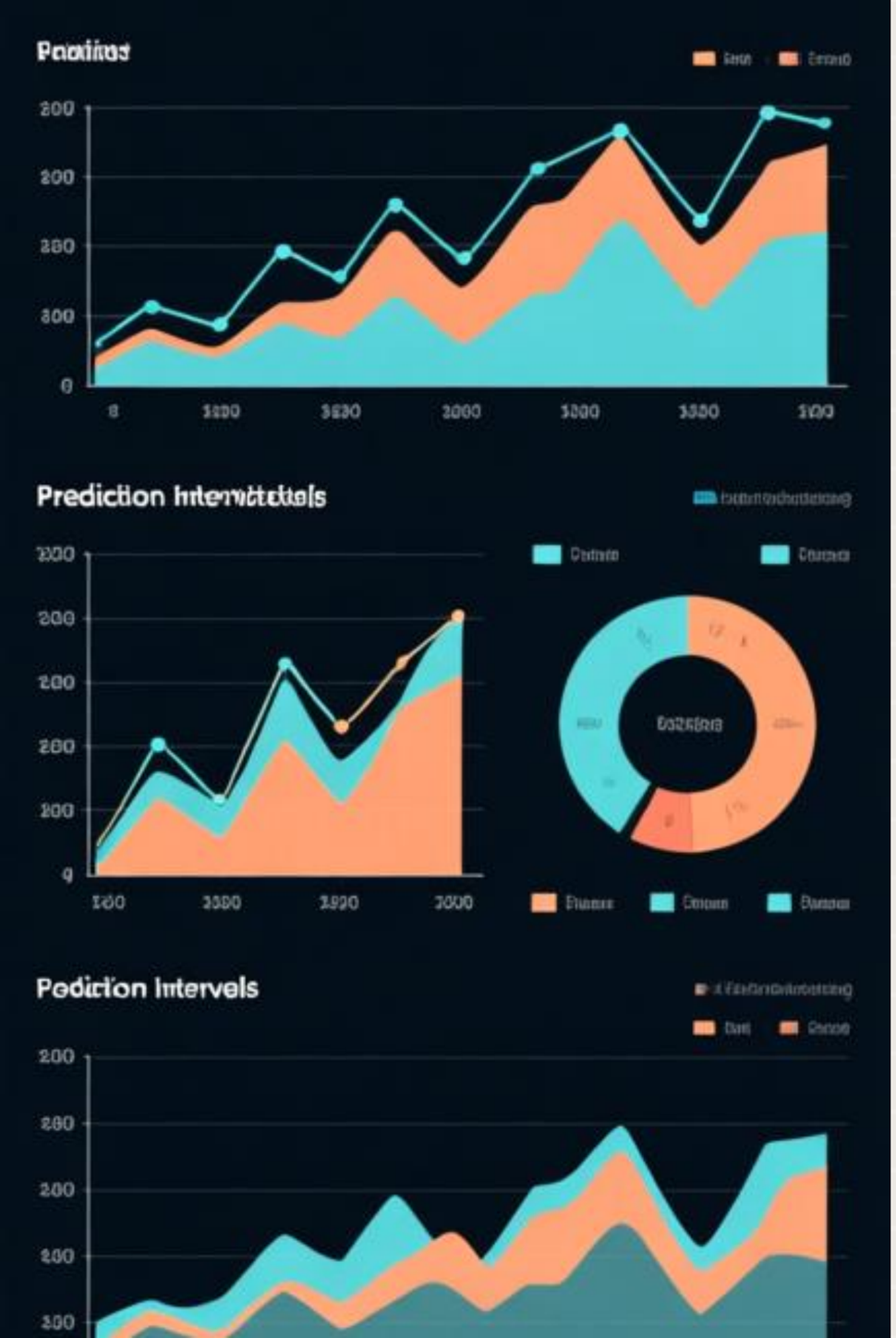
100+ Billion Data Points

100+ Billion Data Points

This massive diversity ensures robust learning across varied temporal patterns, noise types, and outliers.

- Finance
- Economics
- Healthcare
- Weather
- IoT
- Energy
- Web Traffic
- Sales
- Transport
- Banking

TimeGPT trained on 100+ billion data points, the largest publicly available time series collection.



Uncertainty Quantification: Conformal Prediction

TimeGPT uses conformal prediction for probabilistic forecasting. **Generates prediction intervals by performing rolling forecasts on latest available data during inference.** **Provides confidence in predictions for risk-sensitive applications.**

Non-parametric

No distributional assumptions

→

Flexible

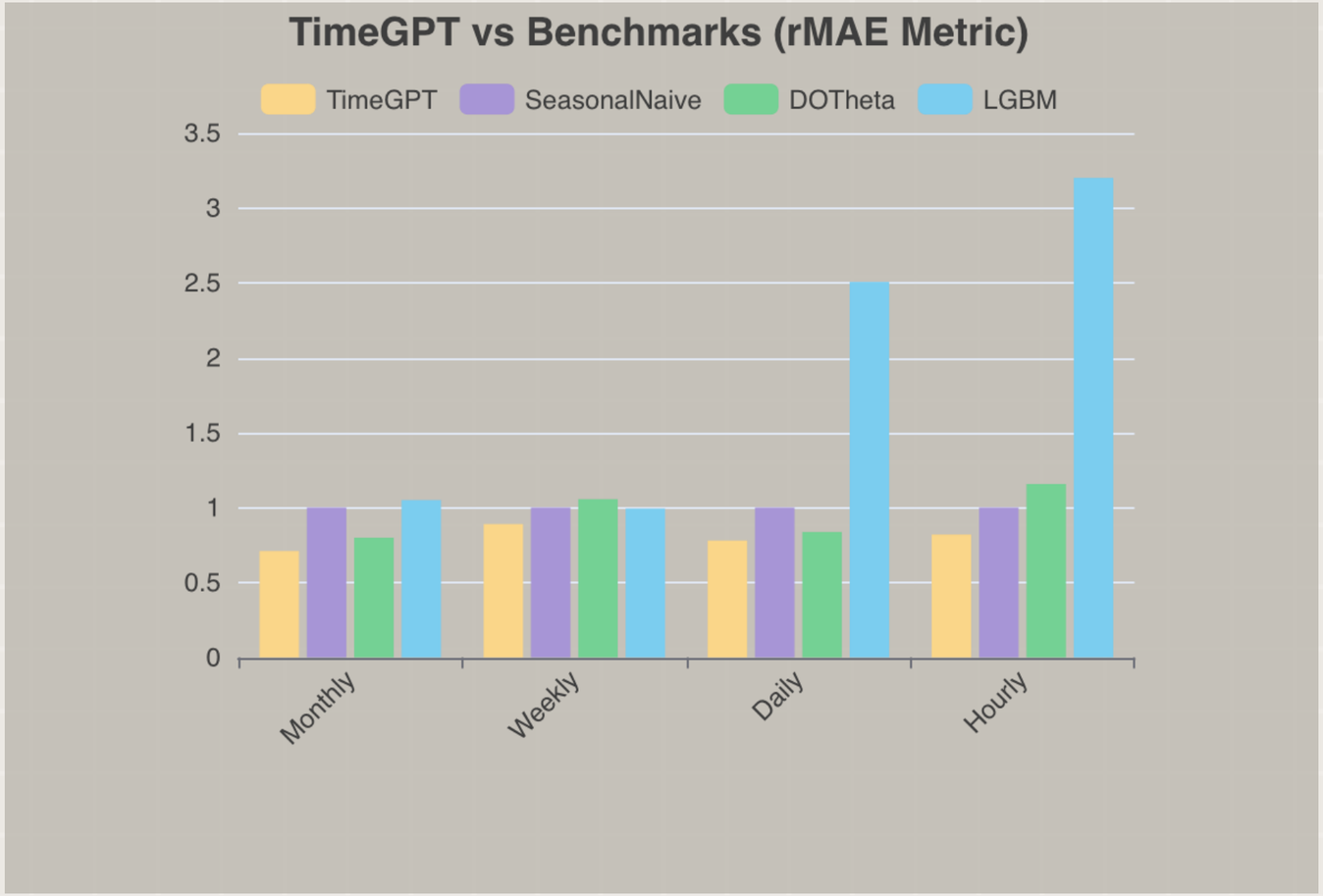
Model-agnostic approach

→

Practical

Pre-specified coverage accuracy

Experimental Results: Zero-Shot Performance Across Frequencies



Superior Performance

TimeGPT ranks top-3 across all frequencies (monthly, weekly, daily, hourly) with zero-shot inference.

Simplicity & Speed

Prediction method is simple and extremely fast, eliminating need for complete training pipelines per dataset.

Fine-Tuning: Tailoring for Domain-Specific Tasks

Foundation Knowledge

Pre-trained model learns from 100B+ data points across domains

Transfer to Target Task

Fine-tuning adapts knowledge to new specific domain

Improved Performance

Domain-specific optimization enhances accuracy (rMAE improvements shown)

Rapid Convergence

Fewer fine-tuning steps needed compared to training from scratch

> "Fine-tuning is critical for foundation models to achieve state-of-the-art domain-specific performance."

Computational Efficiency: Speed Advantage

Scenario 1

TimeGPT Zero-Shot Inference: 0.6 milliseconds per series GPU

Fastest, no training required

Scenario 2

Statistical Methods (Training + Inference): 600 milliseconds per series

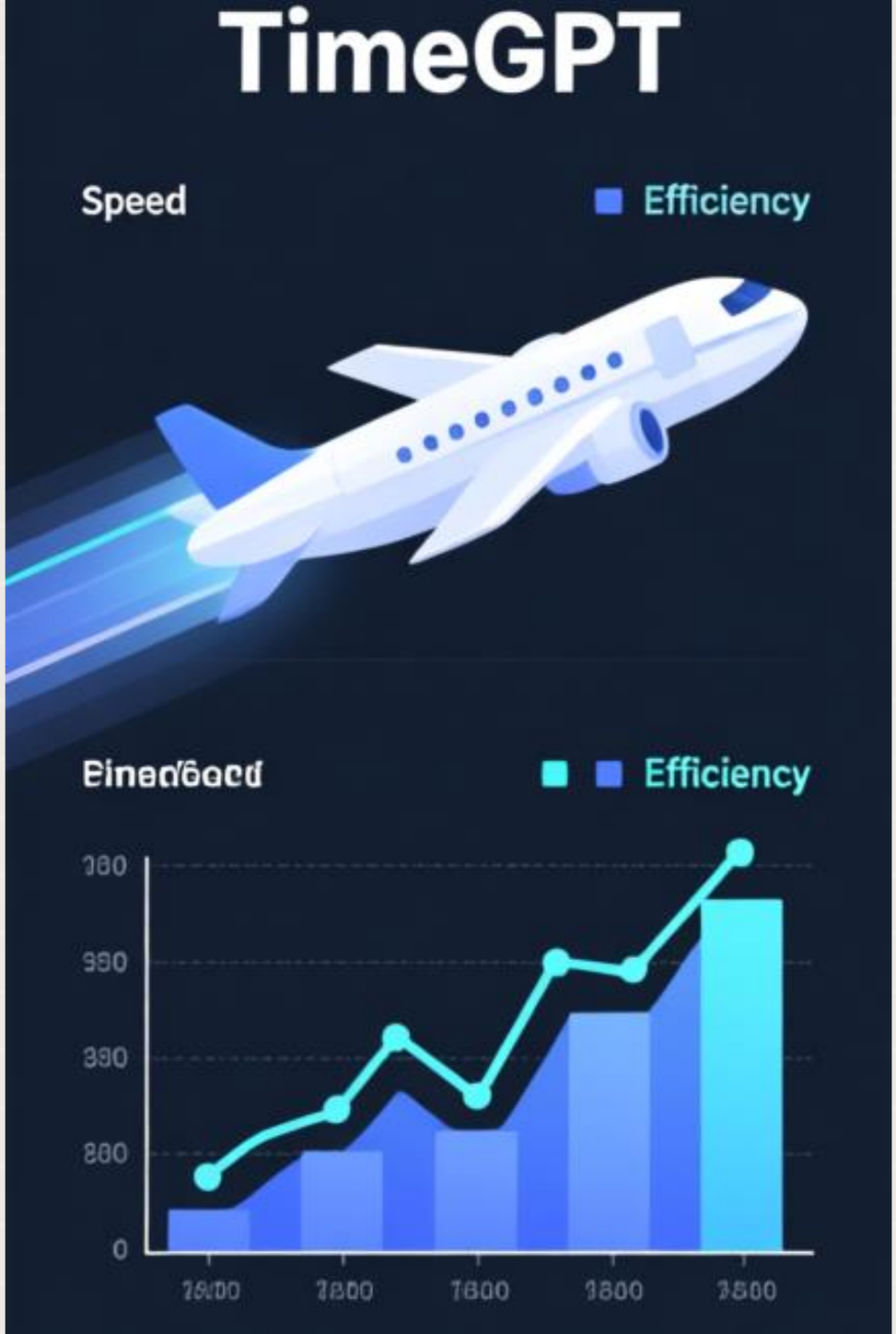
1000x slower than TimeGPT

Scenario 3

Global Models LGBM/LSTM/NHITS (Training + Inference): 57 milliseconds per series

95x slower than TimeGPT

Key insight: "TimeGPT achieves superior performance at unprecedented speed."



Detailed Performance Benchmark: Zero-Shot Inference Results

Comprehensive comparison of TimeGPT against leading forecasting models across multiple frequencies (Monthly, Weekly, Daily, Hourly). Results shown using rMAE and rRMSE metrics, with SeasonalNaive as baseline (1.0).

Model	Monthly rMAE	Weekly rMAE	Daily rMAE	Hourly rMAE
TimeGPT	0.71	0.89	0.78	0.82
DOTheta	0.799	1.056	0.837	1.157
Theta	0.839	1.061	0.841	1.163
SeasonalNaive	1.000	1.000	1.000	1.000
ETS	0.942	1.079	0.944	0.998
CES	1.024	1.002	0.919	0.878
ADIDA	0.852	1.364	0.908	2.307
LGBM	1.050	0.993	2.506	3.200

Key Advantages: Why TimeGPT Stands Out

Simplicity

Reduces complex forecasting pipelines to single inference step

Democratization

Provides access to large transformer models for all

Zero-Shot Performance

Accurate predictions without domain-specific training

Speed & Efficiency

Orders of magnitude faster than alternatives

Robustness

Trained on diverse datasets for generalization

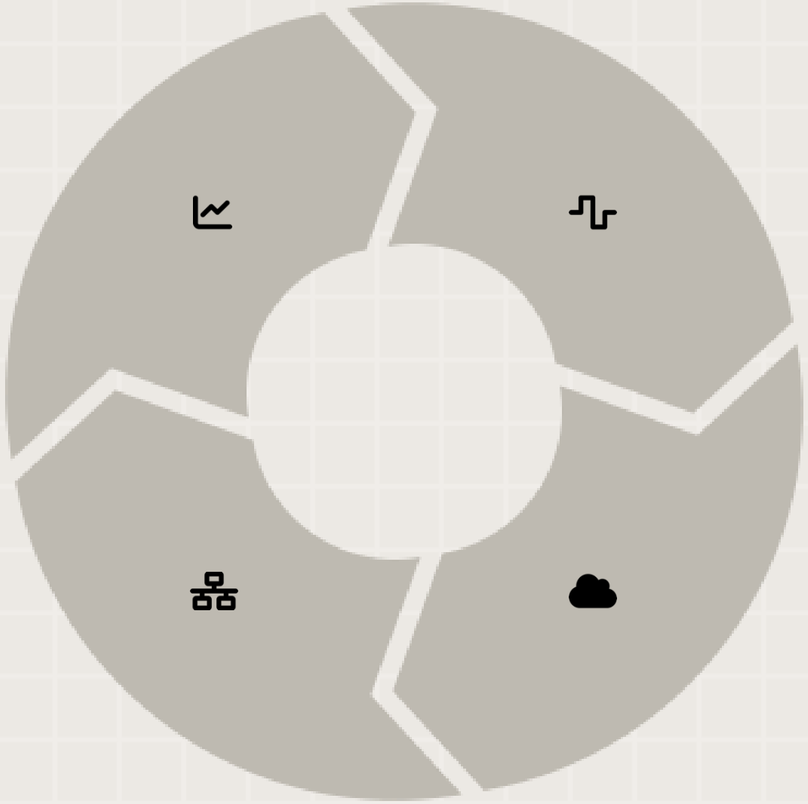
Future Research Directions

Informed Forecasting

Incorporating domain knowledge and expert insights into predictions

Multi-Task Learning

Extending to related forecasting tasks



Time Series Embedding

Measuring similarity between series for transfer learning

Uncertainty Quantification

Enhancing probabilistic prediction methods

Advancing Foundation Models for Time Series

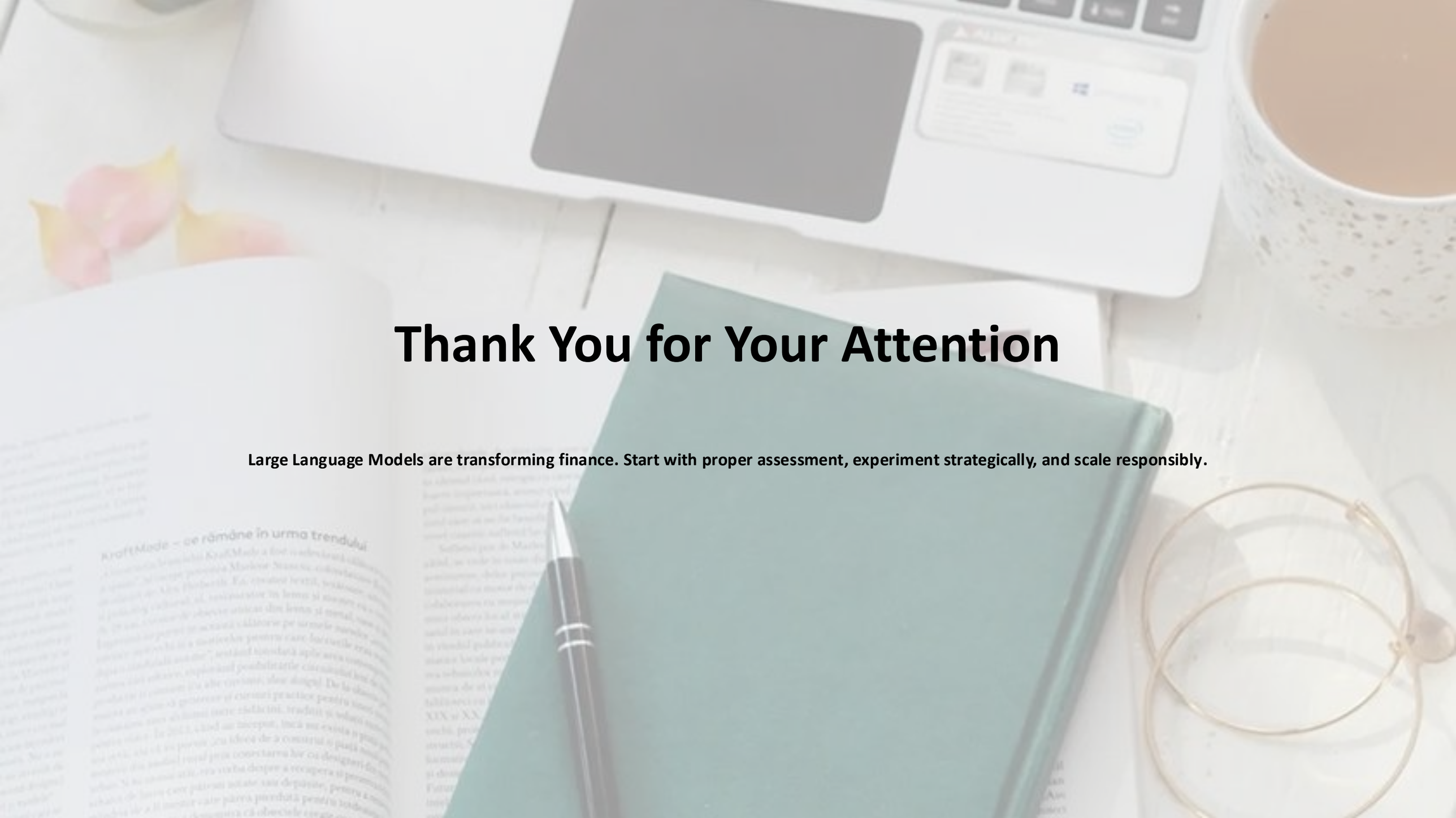
Technical Contributions & Innovation Highlights

Novel Contributions

- First foundation model specifically designed for time series forecasting
- Demonstrates scaling laws benefit time series models
- Zero-shot inference achieves state-of-the-art performance
- Conformal prediction for reliable uncertainty quantification

Innovation Impact

- Eliminates need for dataset-specific model training
- Reduces forecasting complexity by orders of magnitude
- Enables rapid deployment across diverse domains
- Sets new benchmark for foundation model effectiveness



Thank You for Your Attention

Large Language Models are transforming finance. Start with proper assessment, experiment strategically, and scale responsibly.