Learning to Replicate Expert Judgment in Financial Tasks

JUNE 30, 2026

SARAH SU, KEVIN ZHU, EMILY XIAO, ROHAN ALUR, DANIEL KANG
BRIDGEWATER AIA LABS IN COLLABORATION WITH THINKING MACHINES

Investor Judgement

In AIA Labs at Bridgewater, we are building a fully functional artificial investor that can meet or exceed expert human performance across the full range of activities our human investors perform. This is an ambitious goal: beating markets requires differentiated intelligence. When every investor has access to the same sources of information, alpha must come from differentiated judgment and taste. A strong investor's judgment is often difficult to articulate and teach directly to others—whether human or AI—and is typically built up over many years of experience.

To illustrate both the degree of the challenge and an approach to solving it, in this post we consider one of the simplest tasks that our investment analysts routinely perform: filtering and processing financial documents to surface information relevant to investment decisions. Even this tasks turns out to be surprisingly difficult for frontier LLMs.

Investment analysts are bombarded with information every day: news articles, research reports, company documents, emails, internal write-ups, and more. The work isn't the reading itself; it's the small, repeated judgments carried over it: filtering, interpreting, segmenting, and identifying where the useful signal lies. These judgments are embedded throughout an investor's daily workflow and consume substantial time.

We wanted to use AI models to automate the initial step of information triage: identifying what is relevant and interesting to read. This alone could greatly augment analysts' productivity, letting them spend their freed up attention on higher-level synthesis and decision making.

We asked: is it possible to teach LLMs financial taste? We find that **proprietary investor-taste labels**, not merely public financial text, can teach LLMs to interpret text with high taste. **Our trained model outperformed all frontier models on information accuracy and recall, at a fraction of their cost.**

We describe our training process and results on a subset of data that we can make publicly available. This is just one particular illustration of the broader ability that now exists to create *differentiated intelligence*, with models tuned for specific organizational needs. We will share more on how we are pursuing this vision in AIA Labs in subsequent research pieces.

Frontier Model Performance

We evaluated models on six information filtering tasks drawn from investors' daily workflows. Beyond these tasks, we have many others internally that show similar patterns to these six tasks: frontier models we tested on underperform compared to our internally trained models.

We measured accuracy—the percentage of documents that were correctly labeled according to our investors. For classification tasks, we also calculated the F1 score.[1]

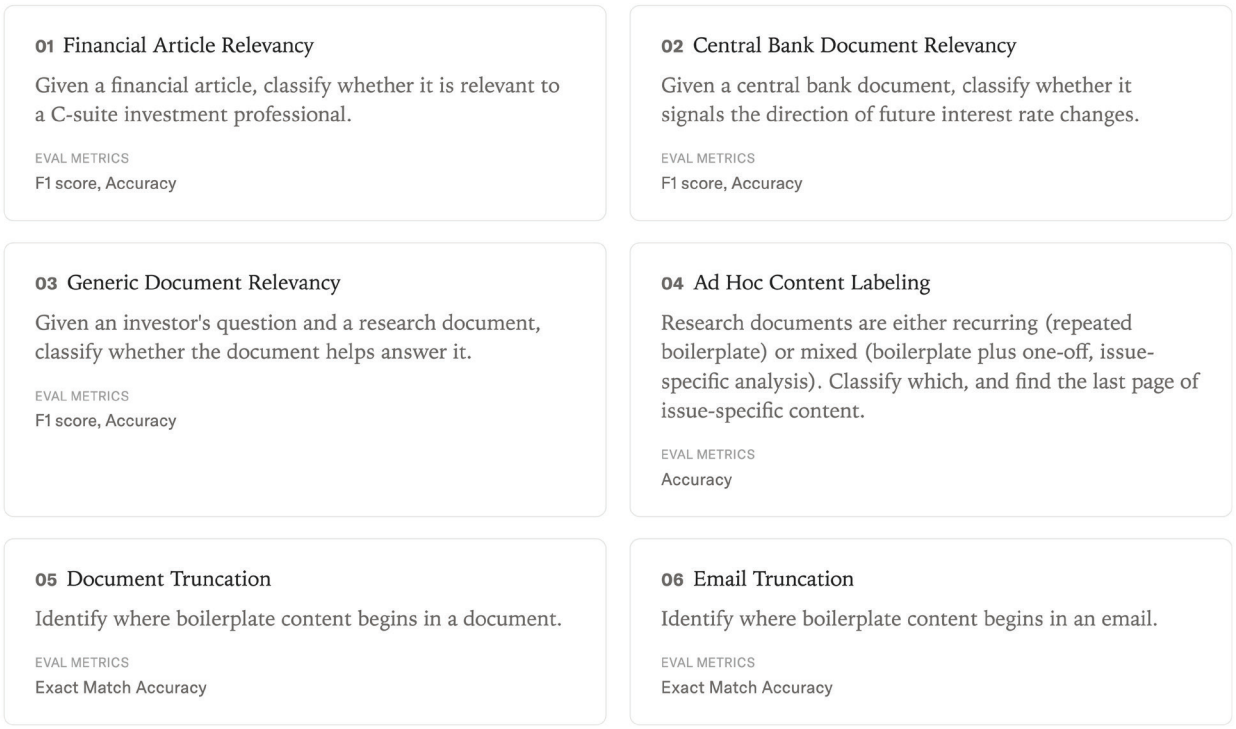

Figure 1: The six financial tasks we evaluate in this blog post, each drawn from the routine work of an investor.

[1] F-score (Wikipedia).

These tasks are trivial for investors, but they get stuck when articulating their decision process. Consider the following example of classifying a news article as relevant to an investment professional below:

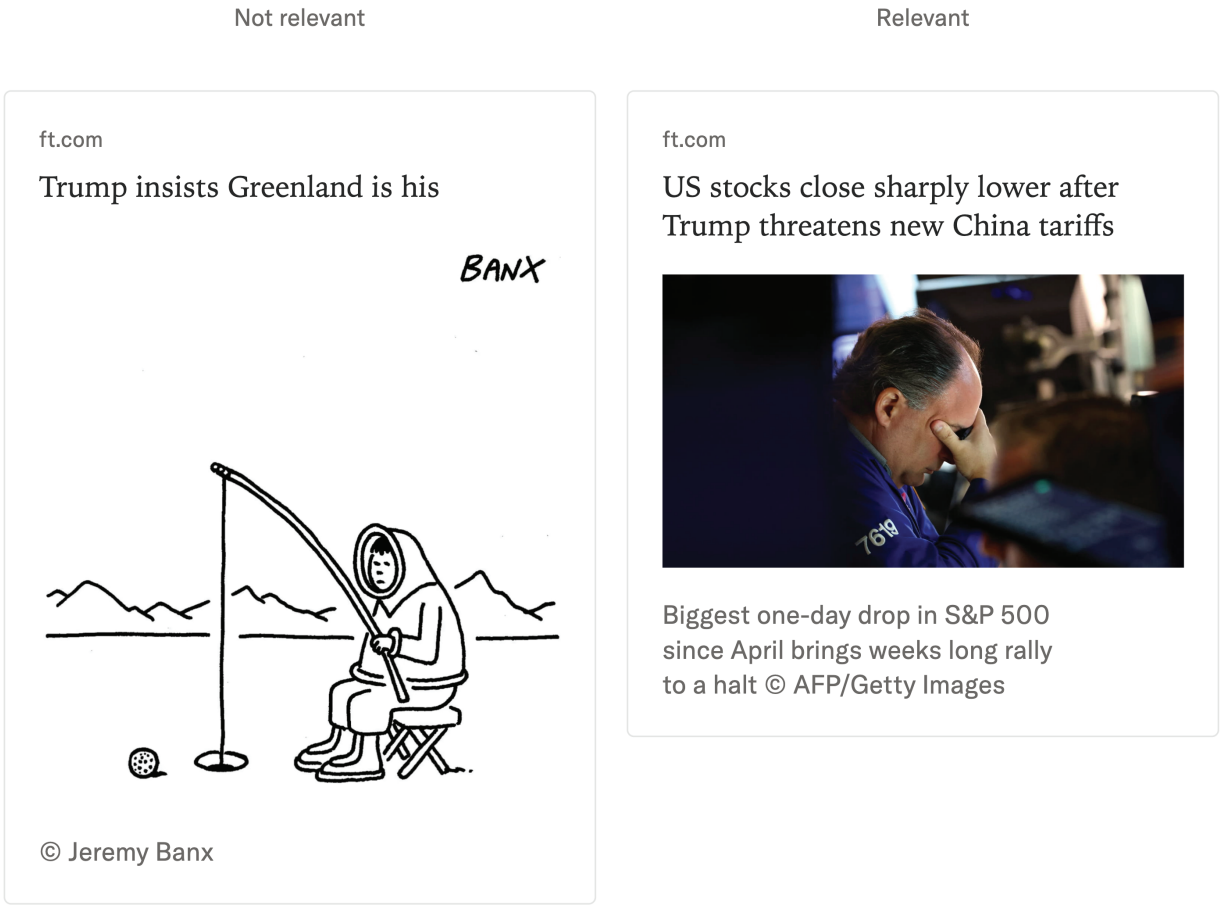

Figure 2: Example of judging the relevance of a financial article to US markets. Source: Financial Times.

The Greenland example is unlikely to be taken seriously given the context of the article, while the China tariffs are highly relevant. Yet both examples touch on geopolitics and finance.

In contrast to our investors, frontier models we tested on perform surprisingly poorly. Variants of Gemini, Claude, and GPT averaged a mere ~50% accuracy when given a prompt that simply states each of the six tasks to perform.

We first tried to improve LLM performance with stronger prompting. Our experts wrote instructions based on real task descriptions, and also suggested reframing certain tasks. For example, while an article about a small IPO is clearly financially relevant, it lacks the broad significance that would make it interesting to a macroeconomic investor at Bridgewater. LLM performance on the article classification task improved when they were asked to sort news stories into three labels: relevant and interesting, relevant but uninteresting, and irrelevant.

These changes boosted their accuracy from a coin flip to the mid-70s. We saw no further gains in accuracy from automatic prompt-optimization methods. With our best prompts the frontier models we tested on still achieved less than 80% accuracy—the threshold investors expect from a system they could trust in their daily workflow.

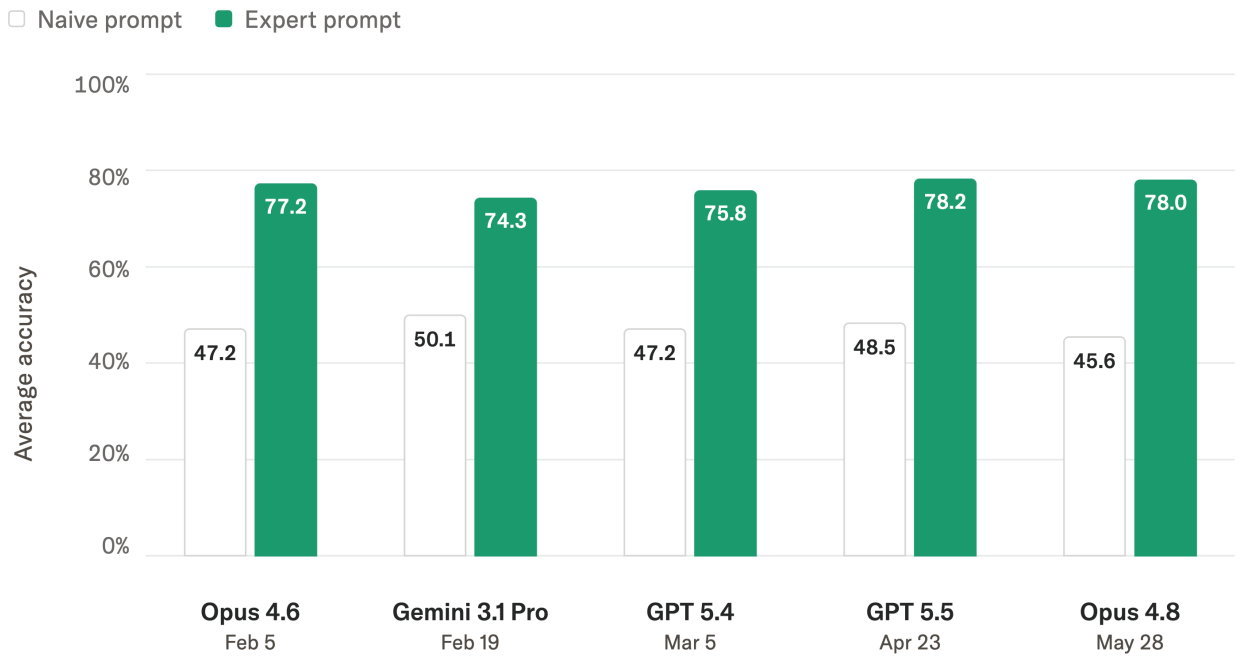

Figure 3: Accuracy & Positive Class F1 score of frontier models on our financial tasks after manual and automatic prompt engineering. F1 score is averaged across our 3 classification tasks, and accuracy is averaged across all 6 tasks.

Our results also suggest that newer models aren't improving rapidly at this task, especially per dollar spent. GPT 5.4 costs 43% more than 5.2 but is only marginally more accurate.

An explicit prompt can only convey the intuition an expert is able to put into words, while the judgments that matter most are often the hardest to articulate. Fine-tuning sidesteps this: rather than contorting the expert's intuition into a static prompt, the training process lets the model develop its own judgment. Could we train open-weight models to outperform frontier models we tested on these tasks?

Training Dataset Construction

The first challenge of training a custom model was acquiring a dataset that reflects **high-quality investor taste**. In particular, much of the information is only useful when filtered through an investment professional's judgment.

We initially sourced a dataset from vendors providing non-expert labeling. Models trained on this dataset still performed poorly. After examining the reasoning traces of the model we realized that the labels in the dataset were often wrong. Since expert labelers are costly, we devised a verification scheme that routes only the contested examples to experts.

The scheme worked as follows: we trained a model on the dataset from non-expert labelers, then evaluated it on the same data. Examples where the model's answer differed from the labelers' were sent to our experts for reevaluation—if a model couldn't match an example from its own training set then either the example is genuinely difficult, or the original label was wrong. This procedure was used to clean the training set data; the final evaluation was done on a held out test set.

Training Recipe

We trained our models on Tinker from Thinking Machines Lab.[2] Tinker allowed us to iterate quickly without worrying about GPU infrastructure.

We chose Qwen3-235B as the base model as its fine-tuning performance is widely studied in the academic literature.

We began with standard GRPO and importance-sampling loss as a simple, critic-free starting point. This baseline approach resulted in a massive jump in the model performance, but it still fell short of our desired 80% threshold.

Model / Training	Average Accuracy	Average Pos F1
Qwen Base	44.8%	55.24%
Qwen + GRPO	73.48%	88.95%

We make the following modifications to our training recipe to push performance farther:

1. ***Interleaved batching***

   For our multi-task training recipe, we compared three batching strategies: training each task sequentially, fully mixing tasks within a batch, and interleaving one batch per task in round-robin order. We found interleaving worked best, improving accuracy by 12.1% over fully mixed batches.

2. ***CISPO loss with asymmetric clipping***

   We used CISPO loss with asymmetric clipping[3] to replace the standard importance-sampling loss. Across the loss functions and clipping schemes we tried, this performed best, improving accuracy by 10.1% over the importance-sampling baseline.

3. ***On-policy distillation with strong teachers***

   We train with on-policy distillation[4] (OPD), constructing the advantage as follows:

$$r = \text{reward} - \beta \cdot \text{avg}(\text{student\_lp} - \text{teacher\_lp})$$

$$\text{adv}_i = r_i - \text{avg}(r)$$

The reward is penalized when the student drifts from the teacher's distribution, regularizing the policy while it learns the task.

Every 20 steps, we promote the current checkpoint to the teacher—but only if validation accuracy has reached a new high, so we never distill toward a weaker model. This gave a further 3.1% gain over a frozen base-model teacher.

[2] Tinker.

[3] CISPO loss with asymmetric clipping (arXiv).

[4] On-Policy Distillation, Kevin Lu in collaboration with others (Thinking Machines).

Results

Finding the optimal training recipe required several iterations of different approaches. Tinker's accessibility allowed us to run fast experiments and refine our approach.

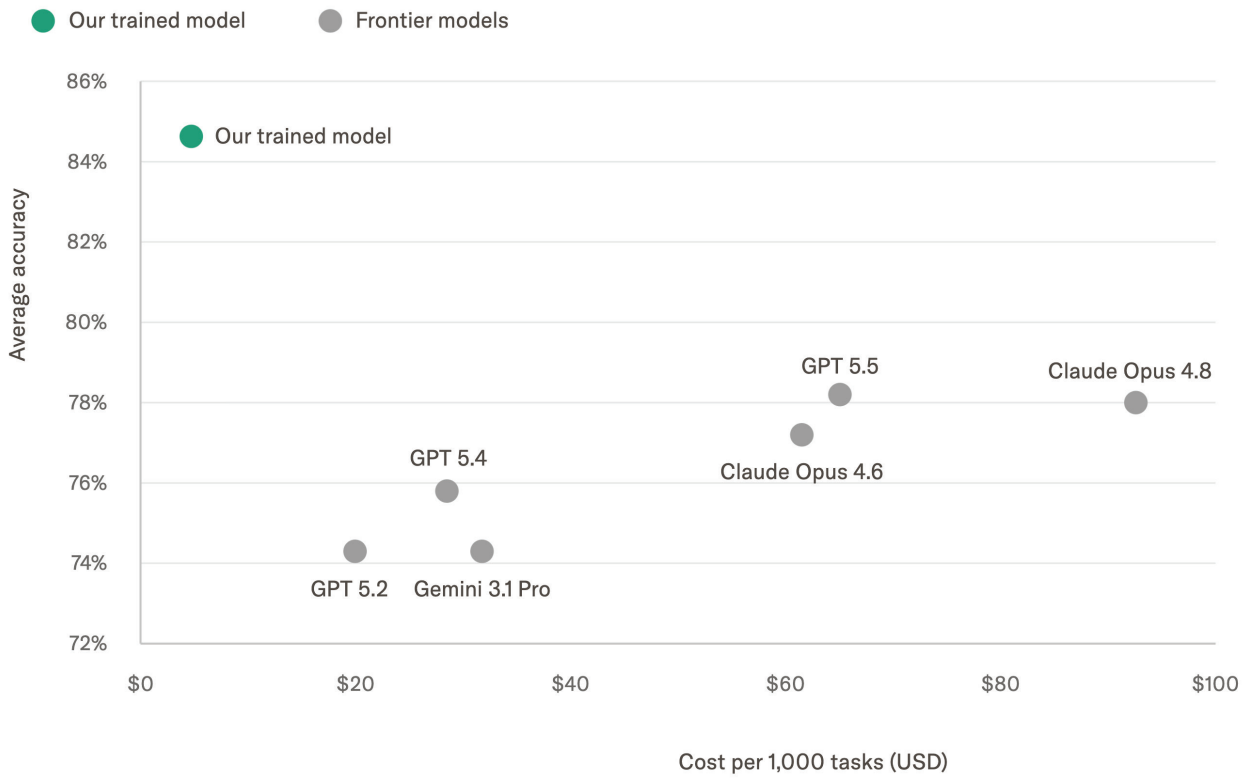

Figure 3: Accuracy versus price for our trained model and frontier models. Our model outperforms frontier models on both dimensions across generations.

Our trained model improves average accuracy from 78.2% to 84.7%, meaning the trained model makes 29.8% fewer mistakes than the best frontier model we evaluated. We find this level of accuracy is sufficient for our daily work.

Our trained model is also vastly cheaper due to its smaller size: a 13.8x reduction in inference costs per task. As we plan to rely on more models trained to help with specific tasks and to scale AI across the organization, cost is an important consideration.

We ablated each part of our training recipe to show how each portion contributes to performance.

Training Method Ablations	Average Accuracy	Avg Pos F1
Qwen + Final Recipe	84.66%	92.99%
⊖ Interleaved Batching	72.18%	89.01%
⊖ CISPO + Asymmetric Clips	74.56%	90.64%
⊖ OPD	72.39%	87.93%
⊖ OPD w/ Best Val Accuracy Teacher	81.55%	89.41%

Figure 4: Each row shows the final recipe with that single component removed (leave one out ablations)

Conclusion

Frontier models struggle with relatively simple financial tasks, and model advances aren't improving performance much. In contrast, we've shown that **high quality proprietary datasets** labeled by expert investors and used for fine-tuning enable custom models that understand our context and perform well on our tasks. We see this result for other tasks and datasets beyond the six we've discussed in the post.

Aside from higher accuracy, custom models are also substantially cheaper. We expect to see more productivity gains from custom model training in the future, especially with the availability of training infrastructure like Tinker that enables rapid experimentation.

Our results show the possibility of a future of differentiated intelligence, where custom models tuned to specific organizational needs outperform frontier models. We'll be releasing more research from Bridgewater AIA Labs in the next few weeks relating to this vision. Stay tuned!

Citation

Please cite this work as:

Su, Sarah; Zhu, Kevin; Xiao, Emily; Alur, Rohan; Kang, Daniel (Bridgewater AIA Labs), "Learning to replicate expert judgment in financial tasks", Thinking Machines Lab: News, June 2026.

Or use the BibTeX citation:

```
@article{su2026expertjudgment,
      author = {Sarah Su, Kevin Zhu, Emily Xiao, Rohan Alur, Daniel Kang (Bridgewater AIA Labs)},
      title = {Learning to replicate expert judgment in financial tasks},
      journal = {Thinking Machines Lab: News},
      year = {2026},
      note = {https://thinkingmachines.ai/news/learning-to-replicate-expert-judgment-in-financial-tasks/}}
```

Important Disclosures and Other Information

This research paper is prepared by and is the property of Bridgewater Associates, LP and is circulated for informational and educational purposes only. There is no consideration given to the specific investment needs, objectives, or tolerances of any of the recipients. Additionally, Bridgewater's actual investment positions may, and often will, vary from its conclusions discussed herein based on any number of factors, such as client investment restrictions, portfolio rebalancing and transactions costs, among others. Recipients should consult their own advisors, including tax advisors, before making any investment decision. This material is for informational and educational purposes only and is not an offer to sell or the solicitation of an offer to buy the securities or other instruments mentioned. Any such offering will be made pursuant to a definitive offering memorandum. This material does not constitute a personal recommendation or take into account the particular investment objectives, financial situations, or needs of individual investors which are necessary considerations before making any investment decision. Investors should consider whether any advice or recommendation in this research is suitable for their particular circumstances and, where appropriate, seek professional advice, including legal, tax, accounting, investment, or other advice. No discussion with respect to specific companies should be considered a recommendation to purchase or sell any particular investment. The companies discussed should not be taken to represent holdings in any Bridgewater strategy. It should not be assumed that any of the companies discussed were or will be profitable, or that recommendations made in the future will be profitable.

The information provided herein is not intended to provide a sufficient basis on which to make an investment decision and investment decisions should not be based on simulated, hypothetical, or illustrative information that have inherent limitations. Unlike an actual performance record simulated or hypothetical results do not represent actual trading or the actual costs of management and may have under or overcompensated for the impact of certain market risk factors. Bridgewater makes no representation that any account will or is likely to achieve returns similar to those shown. The price and value of the investments referred to in this research and the income therefrom may fluctuate. Every investment involves risk and in volatile or uncertain market conditions, significant variations in the value or return on that investment may occur. Investments in hedge funds are complex, speculative and carry a high degree of risk, including the risk of a complete loss of an investor's entire investment. Past performance is not a guide to future performance, future returns are not guaranteed, and a complete loss of original capital may occur. Certain transactions, including those involving leverage, futures, options, and other derivatives, give rise to substantial risk and are not suitable for all investors. Fluctuations in exchange rates could have material adverse effects on the value or price of, or income derived from, certain investments.

Bridgewater research utilizes data and information from public, private, and internal sources, including data from actual Bridgewater trades. Sources include AERIC INC, BCA, Bloomberg Finance L.P., Candeal, Carbon Arc, CEIC Data Company Ltd., Ceras Analytics, China Bull Research, Citibank, Clarus Financial Technology, CLS Processing Solutions, Consensus Economics Inc., Consumer Edge, CRU Group, DTCC Data Repository, Ecoanalitica, Energy Aspects Corp, Enverus, EPFR Global, Eurasia Group, Evercore ISI, FactSet Research Systems, The Financial Times Limited, Finaeon, Inc., FINRA, GaveKal Research Ltd., GlobalSource Partners, Goldman Sachs, Harvard Business Review, Haver Analytics, Inc., IEA, Institutional Shareholder Services (ISS), The Investment Funds Institute of Canada, ICE Derived Data (UK), Investment Company Institute, International Institute of Finance, JP Morgan, JTSA Advisors, LSEG Data and Analytics, MarketAxess, Metals Focus Ltd, MSCI, Inc., National Bureau of Economic Research, Neudata, Organisation for Economic Cooperation and Development, Pensions & Investments Research Center, Pitchbook, Political Alpha, Renaissance Capital Research, Rhodium Group, RP Data, Rubinson Research, Rystad Energy, S&P Global Market Intelligence, Sentix GmbH, SGH Macro, Shanghai Metals Market, Smart Insider Ltd., Swaps Monitor, Tradeweb, United Nations, US Department of Commerce, Visible Alpha, Wells Bay, Wind Financial Information LLC, With Intelligence, Wood Mackenzie Limited, World Bureau of Metal Statistics, World Economic Forum, and YieldBook. While we consider information from external sources to be reliable, we do not assume responsibility for its accuracy. Data leveraged from third-party providers, related to financial and non-financial characteristics, may not be accurate or complete. The data and factors that Bridgewater considers within its research process may change over time.

This information is not directed at or intended for distribution to or use by any person or entity located in any jurisdiction where such distribution, publication, availability, or use would be contrary to applicable law or regulation, or which would subject Bridgewater to any registration or licensing requirements within such jurisdiction. No part of this material may be (i) copied, photocopied, or duplicated in any form by any means or (ii) redistributed without the prior written consent of Bridgewater® Associates, LP.

The views expressed herein are solely those of Bridgewater as of the date of this report and are subject to change without notice. Bridgewater may have a significant financial interest in one or more of the positions and/or securities or derivatives discussed. Those responsible for preparing this report receive compensation based upon various factors, including, among other things, the quality of their work and firm revenues.