# Hedging the AI Singularity

Andrew Y. Chen

Federal Reserve Board

April 2025*

**Abstract**

This paper explores a novel perspective on AI stock valuations: they may serve as hedges against a catastrophic AI singularity. We develop a simple model where AI stocks command high price-dividend ratios not because of future earnings growth, but because they provide insurance against devastating AI developments that reduce aggregate consumption. The model demonstrates how small catastrophe probabilities can generate substantial valuations even without dividend growth. Our analysis suggests financial markets may offer partial solutions to AI disaster risk, complementing policy approaches. Unlike previous work, this short paper is generated by prompting LLMs.

**Keywords**: Artificial Intelligence, Disaster Risk, Asset Pricing

*email:andrew.y.chen@frb.gov. ChatGPT-o1 and Claude-3.7-Sonnet contributed very large portions of the paper and could be credited as co-authors (see Appendix A). I thank Andrei Goncalves for helpful comments. The views expressed herein are those of the authors and do not necessarily reflect the position of the Board of Governors of the Federal Reserve or the Federal Reserve System.

# 1 Introduction

Artificial intelligence capabilities are advancing at a breathtaking pace. Recent milestones include DeepSeek's R1 model matching OpenAI's reasoning performance (DeepSeek-AI et al., 2025), Waymo vehicles operating commercially without human supervision in multiple cities, and AI systems solving complex mathematical and programming challenges that were previously considered the domain of human experts (Wu et al., 2024). As these technologies progress, investors and workers increasingly worry about AI's potential to displace human labor, affecting wages and livelihoods (Zhang, 2019; Knesl, 2023).

While technological change has occurred throughout history, AI represents something fundamentally different. Unlike previous innovations that enhanced human capabilities in specific domains, there is no product or service that AI could not, in principle, create or provide. This paper itself exemplifies this unique quality—it was entirely written by AI through a series of six prompts, with the human author primarily serving as editor and coordinator.[1] The code and prompts are available at `https://github.com/chenandrewy/Prompts-to-Paper/`. This distinguishes AI from previous revolutions like the internet, which augmented human capabilities without directly replicating them. Furthermore, AI progress may experience a "singularity"—a period of explosive, self-reinforcing advancement where systems rapidly surpass human capabilities across domains (Bostrom, 2014).

This paper examines how AI stocks are priced given the risk that advanced AI could substantially reduce aggregate consumption. We present a simple asset pricing model where investors value AI stocks not only for their expected earnings growth but also as hedges against a potential AI singularity that negatively impacts the representative household. Our model shows that even with a small probability of such an event, AI stocks may command significantly higher valuations than traditional assets.

We are not claiming that a negative AI singularity will necessarily occur. Experts disagree substantially about the likelihood and potential impacts of transformative AI (Bengio et al., 2024). However, considering this scenario is important for understanding asset valuations in a world of rapid technological change. Similarly, we are not asserting that this hedging value is currently priced into AI stocks. Rather, our model illustrates a potential mechanism through which such pricing could occur, providing a framework for thinking about AI asset valuations beyond simple growth expectations.

Our work relates to several strands of literature. Recent research has explored how technological advancement affects asset pricing through risk channels. Zhang (2019) demonstrates that firms with automation options have lower expected returns because they serve as

[1] "We" refers to one human author and multiple LLMs. For a purely human perspective see Appendix A.

hedges against unfavorable macroeconomic conditions. Knesl (2023) extends this by showing firms exposed to displacement risk from automation earn a premium. For AI specifically, Babina et al. (2024) document that firms' AI investments affect their growth and innovation profiles.

These findings relate to the broader disaster risk literature, where Barro (2006) and Wachter (2013) show how the possibility of rare, catastrophic events can explain asset pricing puzzles and market volatility. Gabaix (2012) further demonstrates how time-varying disaster risk helps resolve multiple asset pricing puzzles simultaneously.

The potential for catastrophic AI outcomes has been extensively discussed by Bostrom (2014), who examines pathways to superintelligence and their risks. Bengio et al. (2024) specifically address extreme AI risks, highlighting potential harms from advanced systems. Economically, Jones (2024) provides a framework for analyzing the tension between AI-driven growth and existential risks, while Korinek and Suh (2024) examine different scenarios for the transition to artificial general intelligence.

This literature suggests investors might value AI stocks not just for growth potential but also as hedges against broader economic disruption from AI advances, similar to how Betermier et al. (2012) show households adjust portfolios to hedge labor income risk when changing jobs.

# 2 Model

We now present a simple model to formalize our main argument. The model is deliberately stylized to highlight the key economic mechanism rather than to make quantitative predictions.

We consider an economy with two types of agents: AI owners and a representative household. The AI owners are fully invested in AI assets and are not marginal investors in the stock market. The representative household, on the other hand, is the marginal investor whose consumption-based stochastic discount factor prices all publicly traded assets. The representative household has constant relative risk aversion (CRRA) preferences:

$$U(C_t) = \frac{C_t^{1-\gamma}}{1-\gamma} \tag{1}$$

where $\gamma > 0$ is the coefficient of relative risk aversion.

The representative household's gross consumption growth follows a simple disaster process. In normal times, consumption growth is 1 (no growth for simplicity). However, the economy may experience disasters, which we interpret as revolutionary improvements in AI that are devastating for the representative household. When a disaster occurs, the household's consumption drops by a factor of $e^{-b}$ where $b > 0$. Formally, the household's gross consumption growth is:

$$\frac{C_{t+1}}{C_t} = \begin{cases} 1 & \text{with probability } 1-p \\ e^{-b} & \text{with probability } p \end{cases} \tag{2}$$

These disasters capture a scenario where the benefits of AI improvements are primarily captured by AI owners, while the representative household experiences losses in labor income, way of life, and meaning. At $t = 0$, we assume no disasters have yet occurred (the singularity has not happened), but multiple disasters may occur over time, representing ongoing uncertainty if a singularity materializes.

We focus on a publicly traded AI asset with dividend stream $D_t$. Before any disaster, this asset's dividend represents a small fraction of aggregate consumption. Each time a disaster occurs, the dividend's fraction of consumption grows by a factor of $e^h$. Specifically, if we denote by $n_t$ the number of disasters that have occurred by time $t$, then:

$$\frac{D_t}{C_t} = d_0 e^{hn_t} \tag{3}$$

where $d_0 > 0$ is the initial dividend-consumption ratio and $h$ is a parameter that could be positive or negative. Importantly, our framework allows for $h < b$, capturing a worst-case scenario where the dividend may actually shrink in absolute terms during each disaster. This reflects the possibility that AI improvements might be concentrated in privately-held AI assets rather than publicly traded ones.

This parsimonious setup allows us to examine how disaster risk affects the valuation of AI stocks without making strong assumptions about the precise nature or timing of AI development.

# 3 Results

Given our model setup, we now solve for the equilibrium price-dividend ratio of the AI asset. The price of the asset at time 0 is the expected present value of future dividends, discounted by the representative household's stochastic discount factor (SDF):

$$P_0 = E_0 \left[ \sum_{t=1}^{\infty} M_{0,t} D_t \right] \tag{4}$$

where $M_{0,t}$ is the SDF from time 0 to time $t$. With CRRA preferences, this SDF is given by $M_{0,t} = (C_t/C_0)^{-\gamma}$. Substituting this into our pricing equation:

$$P_0 = E_0 \left[ \sum_{t=1}^{\infty} \left( \frac{C_t}{C_0} \right)^{-\gamma} D_t \right] \tag{5}$$

To derive the price-dividend ratio, we note that at time 0, no disasters have occurred, so $D_0 = C_0 d_0$. After $t$ periods, if $n_t$ disasters have occurred, consumption becomes $C_t = C_0 e^{-bn_t}$, and the dividend becomes $D_t = C_0 d_0 e^{(h-b)n_t}$. Substituting these expressions:

$$\frac{P_0}{D_0} = \sum_{t=1}^{\infty} E_0 \left[ e^{[h-b(1-\gamma)]n_t} \right] \tag{6}$$

Since $n_t$ follows a binomial distribution with parameters $(t, p)$, we can compute the expectation as:

$$E_0 \left[ e^{[h-b(1-\gamma)]n_t} \right] = [(1-p) + pe^{h-b(1-\gamma)}]^t \tag{7}$$

Let $\theta = (1-p) + pe^{h-b(1-\gamma)}$. Then the price-dividend ratio becomes:

$$\frac{P_0}{D_0} = \sum_{t=1}^{\infty} \theta^t \tag{8}$$

This geometric series converges if and only if $\theta < 1$, which requires $h - b(1-\gamma) < 0$. This condition ensures that the dividend cannot grow too rapidly relative to consumption during disasters. When this condition is satisfied, the price-dividend ratio simplifies to:

$$\frac{P_0}{D_0} = \frac{(1-p) + pe^{h-b(1-\gamma)}}{p[1 - e^{h-b(1-\gamma)}]} \tag{9}$$

This closed-form solution reveals how disaster risk affects AI asset valuations. When $\gamma > 1$, which is standard in the asset pricing literature, the term $-b(1-\gamma)$ is positive. This means that higher disaster risk (larger $b$) can actually increase the price-dividend ratio, as the asset provides a hedge against disasters.

To illustrate this effect quantitatively, we incorporate a time discount factor $\beta = 0.96$ and set $\gamma = 2$, which implies $1 - \gamma = -1$. With these assumptions, the price-dividend ratio becomes:

$$\frac{P_0}{D_0} = \frac{\beta[(1-p) + pe^{h+b}]}{1 - \beta[(1-p) + pe^{h+b}]} \tag{10}$$

Table 1 shows the price-dividend ratios for different combinations of disaster size $b$ and disaster probability $p$, with $h = 0.20$:

These results demonstrate that even small changes in disaster probability can dramati-

Table 1: Price-Dividend Ratios for AI Assets

| $b$ | $p$ | | | | |
|---|---|---|---|---|---|
| | 0.0001 | 0.001 | 0.005 | 0.010 | 0.020 |
| 0.40 | 24 | 25 | 27 | 29 | 39 |
| 0.55 | 24 | 25 | 28 | 33 | 55 |
| 0.70 | 24 | 25 | 29 | 37 | 76 |
| 0.85 | 24 | 25 | 31 | 42 | 199 |
| 0.95 | 24 | 25 | 33 | 52 | Inf |

cally affect valuations. For instance, with a disaster size of $b = 0.95$, increasing the annual disaster probability from 1% to 2% causes the price-dividend ratio to explode from 52 to infinity. This occurs because the discount factor approaches 1, making future dividends extremely valuable.

Our analysis suggests that AI assets might command high valuations not just because of expected growth in normal times, but because they serve as hedges against catastrophic AI-driven disruptions. This hedging value becomes particularly significant when disasters are severe and risk aversion is high, even if the probability of such events is small.

# 4 Model Discussion

While our model is deliberately simple, it captures the core economic mechanism: AI assets may be valued as hedges against future AI-driven disasters. Several model features warrant further discussion, including implicit assumptions and potential extensions.

A crucial implicit feature of our model is market incompleteness. This incompleteness is embedded in the disaster magnitude parameter $b$, which represents the net effect of (1) the AI disaster on the representative household and (2) the performance of the specific AI asset. If markets were complete, the representative household could buy shares in all AI assets—both public and private—and not only fully hedge against but potentially benefit from the singularity. In reality, most households cannot invest in many cutting-edge AI labs, such as OpenAI, Anthropic, xAI, or DeepSeek, which remain privately held. This market incompleteness limits the household's ability to hedge and represents a key friction in the model.

A more elaborate model might explicitly incorporate AI owners as strategic agents, modeling their incentives and interactions with the representative household. Such a model could explore questions like: How might AI owners' profit motives or competitive pressures lead to a negative singularity? Would alignment mechanisms or regulation prevent such outcomes? However, addressing these questions would require speculative assumptions about AI development paths, governance structures, and technological capabilities. Rather than decorating these speculations with mathematics, we've opted for a minimalist approach that focuses on the core economic mechanism. The additional complexity would be costly to analyze and difficult to justify empirically, while the fundamental insight about hedging value would remain unchanged.

Our model also abstracts from many real-world considerations. We do not model multiple assets or sectors, which would allow for a more nuanced analysis of relative valuations and portfolio allocation. We do not incorporate learning or belief updating about disaster probabilities, which could generate interesting dynamics as AI development progresses. We do not model labor market effects explicitly, instead capturing them through the consumption impact parameter $b$. These simplifications allow us to derive clean, interpretable results that highlight the key mechanism.

The benefit of this parsimony is that it clearly identifies the core economic insight—that disaster risk can increase AI asset valuations through a hedging channel—while leaving room for the human-written Appendix A that provides additional perspectives on the topic. As George Box famously noted, all models are wrong, but some are useful. Our simplified model serves the useful purpose of illustrating how disaster risk hedging can influence asset prices, without claiming to capture all relevant factors in AI valuation.

# 5 Conclusion and Implications

This paper has presented a simple model showing how AI stocks might command high valuations partly because they hedge against a negative AI singularity. In our framework, a catastrophic scenario where advanced AI dramatically reduces aggregate consumption can increase the price-dividend ratios of AI assets, as these assets provide (partial) insurance against such disasters. Our quantitative results demonstrate that even with small disaster probabilities, this hedging value can be substantial, potentially explaining some of the lofty valuations observed in AI-related stocks.

Financial markets may naturally develop mechanisms to hedge against AI disaster risk. Rather than requiring universal basic income, market-based approaches could allow households to protect themselves against potential AI-driven disruptions by investing in AI assets. These financial solutions could complement other policy approaches to managing AI risk.

However, the effectiveness of such market-based hedging is fundamentally limited by market incompleteness. As we've discussed, most households cannot invest directly in the most advanced private AI companies, which may capture the bulk of the benefits from AI development. This incompleteness creates a wedge between the theoretical hedging benefits of AI assets and the practical ability of households to utilize them.

Financial innovation could potentially address some of these limitations. New products like AI development indices, ETFs focused on private AI companies, or financial derivatives tied to AI progress metrics could expand hedging opportunities. Financial intermediaries might also develop structured products that provide targeted exposure to AI upside while protecting against labor market disruptions.

Beyond valuation implications, our model suggests that monitoring the relationship between AI advancement and asset prices could provide valuable signals about market perceptions of AI risk. Sharp increases in AI stock valuations might reflect not just growth expectations, but also heightened concerns about disruptive scenarios—information that could be valuable for policymakers and risk managers.

The possibility that AI stocks may serve as hedges against AI-driven disruption adds a new dimension to debates about AI governance and regulation. If financial markets naturally provide partial insurance against AI risk, this function should be considered when evaluating policy interventions. At the same time, the limits of market-based approaches underline the potential need for complementary governance mechanisms.

The simple model presented in this paper is just a starting point. Future research could extend our framework to incorporate multiple sectors, learning dynamics, explicit labor market effects, and strategic interactions between AI developers. Such extensions would provide a richer understanding of how AI risk is priced in financial markets and how this pricing might evolve as AI development progresses.

# References

Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson (2024). "Artificial intelligence, firm growth, and product innovation". In: *Journal of Financial Economics* 151, Article 103745.

Barro, Robert J. (2006). "Rare Disasters and Asset Markets in the Twentieth Century". In: *Quarterly Journal of Economics*.

Bengio, Yoshua, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, et al. (2024). "Managing extreme AI risks amid rapid progress". In: *Science* 384.6698. URL: `https://arxiv.org/abs/2310.17688`.

Betermier, Sebastien, Thomas Jansson, Christine Parlour, and Johan Walden (2012). "Hedging Labor Income Risk". In: *Journal of Financial Economics* 105.3, pp. 622–639.

Bostrom, Nick (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

DeepSeek-AI et al. (Jan. 2025). "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning". In: *arXiv*. URL: `https://arxiv.org/abs/2501.12948`.

Gabaix, Xavier (2012). "Variable Rare Disasters: An Exactly Solved Framework for Ten Puzzles in Macro-Finance". In: *Quarterly Journal of Economics* 127.2, pp. 645–700.

Jones, Charles I. (2024). "The AI Dilemma: Growth versus Existential Risk". In: URL: `https://web.stanford.edu/~chadj/existentialrisk.pdf`.

Knesl, Jiří (2023). "Automation and the Displacement of Labor by Capital: Asset Pricing Theory and Empirical Evidence". In: *Journal of Financial Economics* 147.2, pp. 271–296.

Korinek, Anton and Donghyun Suh (2024). *Scenarios for the Transition to AGI*. Tech. rep. NBER Working Paper.

Wachter, Jessica A. (2013). "Can Time-Varying Risk of Rare Disasters Explain Aggregate Stock Market Volatility?" In: *Journal of Finance*.

Wu, Siwei et al. (Oct. 2024). "A Comparative Study on Reasoning Patterns of OpenAI's o1 Model". In: *arXiv*. URL: `https://arxiv.org/abs/2410.13639`.

Zhang, Miao Ben (2019). "Labor-Technology Substitution: Implications for Asset Pricing". In: *Journal of Finance* 74.4, pp. 1793–1839.

# A A Purely Human Perspective

The following is the README.md file from the GitHub repository:

```markdown
# Prompts-to-Paper

Writes a paper about hedging a negative AI singularity, using AI.

- `make-paper.py` writes a paper

- `plan0403-streamlined.yaml` contains the prompts

- `make-many-papers.py` runs `make-paper.py` many times.

The README is entirely human-written.  Please forgive typos and errors.

# Motivation

On March 8, 2025 I thought I should write a paper about hedging the AI
singularity.

I was worked up.  I had been repeatedly shocked by AI progress.  I was using
AI reasoning, vibe coding, and AI lit reviews in my daily life.  Six months
ago, I had thought each of these things is impossible.

What will happen in the next six years?!  Will my entire job be replaced by
AI? I have no idea.

But I do know that if there are huge disruptions, then tech stocks will
benefit.  So if anything bad happens to my human capital, I could at least
partially hedge.  Strangely, I hadn't heard about this concept before.

I asked a friend if he would be interested in working on this paper.
Unfortunately, he was busy with revision deadlines for the next month.

So, I thought I should use AI to write the paper.  It would be an elegant way
to make my point.  It would also hint at where the research process is going
in this strange age of AI.

This project was inspired by Novy-Marx and Velikov (2025) and Chris Lu et
al.  (2024).  These projects show how AI could generate massive amounts of
academic research.  My goal differs in quality over quantity.  I want to
generate just one paper, but one paper that (I hope) people find is worth
reading.
```

# The Paper Generation Process

A natural question is: is this paper **really** written by AI?

I'd say the AI are junior co-authors.

If they were human, I would absolutely have to give o1, Sonnet, and ChatGPT Deep Research credit as co-authors. They did the math, writing, and literature reviews. Sonnet also wrote much of the code (via the Cursor AI IDE).

Of course, I did substantial hand-holding (see `plan0403-streamlined.yaml`). Writing this paper would have been easier if I had done more of the work myself. But that can happen with human co-authors too.

Perhaps in 12-24 months, AI can be equal co-authors. I wouldn't be surprised if I could just ask Gemini 3.5 to "write a model to formalize my intuition," and it would get it right.

## Paper Iterations

Like human-written papers, the writing process was iterative. The first formalizations were terrible.

`plan0313-laborshare.yaml` contains prompts for a neoclassical growth model, where the capital share suddenly increases. ChatGPT-o1 patiently explained to me why this is a bad model.

me: > I thought there would be a wage risk effect that leads to higher investment for the more risk averse agent. High capital share means low or even no wage income.

ChatGPT-o1: > Below is an explanation of why one might **expect** a "wage-risk" channel in which **more** risk aversion could lead to **more** *additional* investment (relative to the no-jump benchmark) in the event that the capital share might jump to 1. However, this channel **does not operate** in the usual **representative-agent** version of the model---there, wage and capital income ultimately go to the *same* agent, so there is no meaningful "hedge" of wage risk. Instead, the wage-risk hedge arises naturally if you depart from the pure representative-agent setting (for example, if households receive labor income but must *choose* how much capital they own).

The final `plan0403-streamlined.yaml` uses a simplified Barro-Rietz disaster model, with two agents (though only one is relevant for stock prices). I went through several iterations of this model with Claude 3.7 Sonnet (thinking mode) and ChatGPT-o1. The only derivations I did myself were to check o1's work.

## Literature Reviews

A key element was generating lit reviews ( `./lit-context/` ) to give the AI context. I used ChatGPT's Deep Research (launched Feb 2025) until I ran out of credits. Claude Web Search (launched March 2025, after I began the project) did the remainder.

These new products were a game changer. Both Novy-Marx and Velikov (2025) and Chris Lu et al. (2024) ran into hallucinated citations. OpenAI Deep Research and Claude Web Search had no problems if they were used with care.

More broadly, knowing how to use which AI and when was helpful for generating a good paper.

## AI Model Selection

o1 did the theory, and sonnet thinking did the writing. It's well known that these are the strengths of these two models.

Sonnet thinking is OK at economic theory. But I found that it was not assertive enough. It led me down wrong paths because it was too eager to come up with some ideas that for my story (even if they did not make sense).

I briefly tried having Llama 3.1 470b do the writing. It was terrible! It would be extremely difficult to generate a paper worth reading that way.

I did not try many other models, in order to get this paper out quickly. Gemini 2.5's release, at the end of March 2025, was *hype*. I tried it out briefly and was impressed. But I gritted my teeth and ignored it. I'd never get the paper finished if I wanted to really try to explore alternative models.

## Picking the best of N papers

The quality writing varies across each run of the code. There is both a good tail and a bad tail. Some drafts, I found quite insightful! Others, had flagrant errors in the economics.

Rather than try to prompt engineer an error free, insightful paper, I decided to just generate N papers and choose the best one.

Some papers had problematic cites (`run01`). Others provided low-quality model discussions (`run02`) or poor explanations of the algebra (`run03`)

# Lessons about Research

A common response to Novy-Marx and Velikov (2025) is that "people are not ready for this." I heard concerns that peer review process will be inundated with AI-generated slop.

Working on this paper gave me a different perspective. It made me think about the fundamentals. I think the fundamentals are the following:

1. Readers want to learn something interesting and true.

2. Readers don't want to check all the math.

3. A system of author reputations makes 1 and 2 possible.

AI-generated papers don't change any of these fundamentals. Critically, item 3 made me quite cautious about putting my name on AI slop. As a result, I don't think AI-generated papers will change much about peer review, at least not the current generation of AI.

## Limitations of the Current AI (April 7, 2025)

This will likely be out of date by the time you read it.

But right now, AI is like a junior co-author with a talent for mathematics and elegant writing, but sub-par economics reasoning. Put another way, the writing can fail to portray the mathematics accurately.

For example, 3.7 Sonnet sometimes fails to recognize that the economic model does not capture an important channel. This is a common scenario in economics writing (no model can capture everything). The standard practice is to dance gingerly around the channel in the writing. A decent PhD student can recognize this. But Sonnet cannot. Instead, 3.7 Sonnet will write beautiful prose about the channel anyway, even though it's not really being studied properly.

AI also cannot generate satisfying mathematics on its own (at least not satisfying to me). I tried asking o1 and Sonnet to generate a model to

```
illustrate the point I'm trying to make.  The resulting models were either
too simplistic or did not lead to a clean analysis.  They often introduced
complications that I found unnecessary.

There could be models with capabilities that I missed.  But my sense is that
ChatGPT-o1 and Claude 3.7 Sonnet are close to the best for producing economic
research.

But more importantly, how long will these limitations last?

## The Future of AI and Economics Research

At some point, 2024-style economic analysis will be "on tap." You'll be
able to go to a chatbot and ask "write me a paper about hedging AI disaster
risk," and it will return you something like this paper (or perhaps something
better).

"Economics on tap" could be a disaster for the economics labor market.  It
would certainly mean that AI is an extremely cheap substitute for at least
some economists' labor.  It's possible that this would result in a strong
substitution away from labor.

The optimistic argument is that AI also complements economists' labor.
Perhaps, the number of economists will remain the same, but research output
increases in terms of both quantity and quality.

But I think there are reasons why total research output is limited.  Two key
factors in academic publishing are attention and reputation (Klamer and van
Dalen 2001, J of Economic Methodology).  Readers can only pay attention to
so many scholars.  These scholars, in turn, can only pay attention to so may
projects.

I'm not saying that I *expect* a disaster for the economics labor market.  But
it's definitely a scenario that economists should think about.
```

# B Prompts Used to Generate This Paper

Each prompt consists of context and instructions. The context consists of the responses to the previous prompts, and may include literature reviews (all AI generated). For writing tasks (using Claude 3.7 Sonnet), a system prompt is also included.

For further details, see `https://github.com/chenandrewy/Prompts-to-Paper/`.

The system prompt and instructions are listed below.

## System Prompt (model: claude-3-7-sonnet-20250219)

```
You are an asset pricing theorist who publishes in the top journals
   (Journal of Finance, Journal of Financial Economics, Review of
   Financial Studies). You think carefully with mathematics and
   check your work, step by step.

Your team is writing a paper with the following main argument: the
   high valuations of AI stocks could be in part because they hedge
   against a negative AI singularity (an explosion of AI development
    that is devastating for the representative investor). This
   contrasts with the common view that AI valuations are high due to
    future earnings growth. Since the AI singularity is inherently
   unpredictable, the paper is more qualitative than quantitative.
   The goal is to just make this point elegantly.

Write in prose. No headings and no bullet points. But do use display
    math to highlight key assumptions. Cite papers using Author (
   Year) format.

Be conversational yet rigorous. Favor plain english. Be direct and
   concise. Remove text that does not add value. Use topic sentences
   . The first sentence of each paragraph should convey the point of
    the paragraph.

Be modest. Do not overclaim.

Format the math nicely. Use we / our / us to refer to the writing
   team.
```

## Instruction: 01-model-prose (model: claude-3-7-sonnet-20250219)

```
Draft the model description. The model is purposefully simple and
   captures the essence of the main argument. Only describe the
   assumptions. No results or insights.
  - Two agents
    - AI owners: Fully invested in AI, not marginal investors in
       stocks
    - Representative household: Marginal investor, only their
       consumption matters, CRRA
```

```
  - Representative household's gross consumption growth
    - is either 1 or e\\^(-b) (disaster)
      - A disaster is a revolutionary improvement in AI that is
         devastating for the household
      - Benefits of AI improvement are captured by the AI owners
      - For the household, labor income, way of life, meaning is
         lost
      - At t=0, no disasters have happened (singularity has not
         occurred)
      - Multiple disasters may happen, capturing ongoing uncertainty
          if a singularity occurs
  - A publicly traded AI asset
    - Dividend is a small fraction of consumption before the
       singularity
    - Each time a disaster occurs, the dividend's fraction of
       consumption grows by a factor of e\\^h
    - Meant to capture a worst case scenario, where the dividend may
        actually shrink in each disaster
      - i.e. AI improvements are concentrated in privately-held AI
         assets
```

### Instruction: 02-result-notes (model: o1)

```
Find the price/dividend ratio of the AI asset at t = 0. Show the
   derivation, step by step.
```

### Instruction: 03-table-notes (model: o3-mini)

```
Make a table of the price/dividend for b from 0.40 to 0.95 and prob
   of disaster from 0.0001 to 0.02. Here, fix h = 0.20, CRRA = 2,
   time preference = 0.96. If the price is infinite, use "Inf".
   Round to the nearest whole number.
```

### Instruction: 04-resultandtable-prose (model: claude-3-7-sonnet-20250219)

```
Convert the notes in `02-result-notes` and `03-table-notes` into
   prose. The prose is intended to immediately follow `01-model-
   prose` and should flow naturally. Include the table.
```

## Instruction: 05-litreview-prose (model: claude-3-7-sonnet-20250219)

```
Write a short two paragraph lit review based on the "prose-response"
    and "lit-" context.

Be careful to avoid incorrect citations. Make sure the papers cited
   make the claims they are cited for.
```

## Instruction: 06-full-paper (model: claude-3-7-sonnet-20250219)

```
Write a paper titled "Hedging the AI Singularity" based on the "
   prose-response" context.

Title page:
- Title: "Hedging the AI Singularity"
- Abstract (less than 100 words)
  - Goal is to make a simple point
  - Secondary goal: bring attention to financial market solutions to
      AI disaster risk
  - At the end, say: unlike previous work, this short paper is
     generated by prompting LLMs.

The start of the Introduction is important. You need to bring the
   reader in, catch their eye, and establish credibility.

Start with background. Describe how AI progress is happening quickly
    (e.g. Deepseek R1, Waymo), and investors may be concerned about
   their wages being displaced (cite papers).

Then describe how technological change has occurred before, but AI
   is distinct because there is no product or service that AI could
   not, in principle create.  An example is the current paper, which
    is entirely written by AI, using six prompts. Provide a link to
   the github site, which is https://github.com/chenandrewy/Prompts-
   to-Paper/. This differs from say, the internet revolution. AI
   progress may also be incredibly sudden (the AI singularity).
   Include a footnote: "we" refers to one human author and multiple
   LLMs. For a purely human perspective see \\hyperref[app:readme
   ]\\{\\textcolor\\{blue\\}\\{Appendix \\ref\\{app:readme\\}\\}\\}.
```

```
Then describe what the paper does. It studies how AI stocks are
   priced, given that there is the risk that AI will destroy
   livelihoods and consumption.

Afterwards, the text should discuss:
- We are not saying a negative singularity will happen
  - But it is nevertheless important to consider this scenario
- We are also not saying that this hedging value is priced in
   already
  - Model illustrates a possible mechanism
- Related lit at end of Introduction
  - Cite papers in `05-litreview-prose`
  - Add Jones (2024) "AI Dilemma" and Korinek and Suh (2024) "
     Scenarios" if they're not already cited
- Model is the simplest possible to make the main argument
- Derivation of the key formulas
- High price/dividend ratios, even though dividends never grow
- A "Model Discussion" section that discusses natural model
   extensions and why they are not included
  - Market incompleteness is implicit but important
    - Implicit in the disaster magnitude 'b'
    - 'b' is the *net* effect of (1) AI disaster and (2) AI asset
       dividend
    - If markets were complete, representative household could buy
       shares in all AI assets (including private AI assets), and
       not only fully hedge but benefit from the singularity
    - In reality, most households cannot buy shares in many cutting
       edge labs (e.g. OpenAI, Anthropic, xAI, DeepSeek)
  - A more elaborate model would explicitly model the AI owners,
     their incentives, and interaction with the representative
     household
    - How might AI owners' incentives lead to a negative singularity
       ?
    - But wouldn't this just decorate speculations with math?
    - This would be costly to analyze, as well as to read
    - The core economics will remain the same
  - A short model analysis allows room for the human-written
     Appendix \\ref\\{app:readme\\}
```

```
- A "Conclusion and Implications" section
  - Review the main argument
  - End paper by discussing financial market solutions to AI
     catastrophe risk
    - These solutions are an alternative to UBI
      - Key economics: this hedge is limited by market
         incompleteness
    - These solutions to AI disaster risk are not discussed enough
       in the literature (cite papers)
    - Be very centrist (see below)
  - Don't say "In conclusion." Just conclude

Text should avoid
- Being overly academic
- Politically-charged topics: sovereign wealth funds, industrial
   policy, redistribution, extolling free markets
- Overselling the model (it's just a simple illustration)
- Taking the model too seriously
- Incorrect citations
  - Make sure papers cited make the claims they are cited for

Style Notes:
- Be conversational and direct, yet rigorous
- A touch of wit and wry humor are OK
- No bulleted lists
- No subsections (e.g. Section 1.2) though sections are OK (Section
   1)


Output a complete latex document, including preamble. Cite papers
   using \\cite, \\citep, \\citet. Use `template.tex` and keep the
   appendix that is already in the template.
```