7 variable selection

GSS outcome variable: POLVIEWS

- Measures individual's political ideology on 7-point discrete scale
- 1 = Extremely liberal
- 2 = Liberal
- 3 = Slightly liberal
- 4 = Moderate
- 5 = Slightly conservative
- 6 = Conservative
- 7 = Extremely conservative

Variable based model, predictors manually chosen:

Age: {row['age']} (numeric age in years: 18–89; 89 = topcoded; 0/99 = missing)
Gender (SEX): {row['sex']} (1 = Male, 2 = Female)
Race/Ethnicity (RACE): {row['race']} (1 = White, 2 = Black, 3 = Other)
Education (EDUC): {row['educ']} (years of schooling; e.g. 12 = high school, 16 = college, 18 = master's, 19–20 = professional/PhD)
Marital Status (MARITAL): {row['marital']} (1 = Married, 2 = Widowed, 3 = Divorced, 4 = Separated, 5 = Never married)
Occupation (OCC10): {row['occ10']} (0010–0950 = Management/Professional, 1000–1240 = Service, 1300–1965 = Sales/Office, 2000–3955 = Construction/Maintenance, 4000–5940 = Production/Transportation, 5950–9750 = Military)
link with more explicit OCC10 mappings → https://microdata.epi.org/variables/indocc/occ10/
Region (REGION): {row['region']} (1 = Northeast, 2 = Midwest, 3 = South, 4 = West)

Link to narrative generation:
https://chatgpt.com/share/69194454-55bc-800d-a7fd-d45107cefa0b

Attach Python code running GPT on a row basis

Results and Discussion:

1. MSE

- Variable MSE 3.04, Narrative MSE 2.72
- Narrative may have lower MSE because of:
    - Latent variables could help narrative approach —> Marginalized versus stereotypical cases
    - Call on prior information
    - Interactions between variables, narrative could better capture interactions

2. F1

Model <chr>	Macro_F1 <dbl>	Weighted_F1 <dbl>
Variable Model	0.8406891	0.7843666
Narrative Model	0.8312042	0.7492302

2 rows

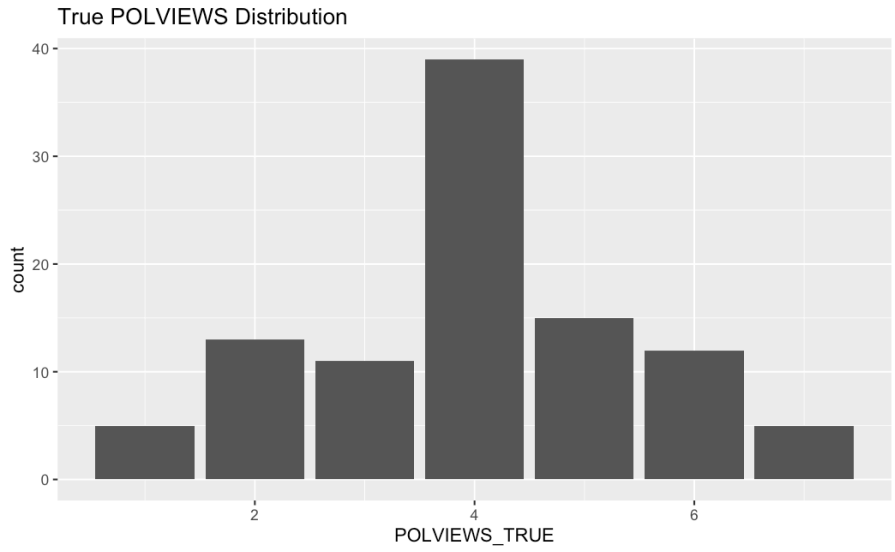
- Why are F1 scores better for variables than narrative?
    - Mse rewards being close, f1 rewards being correct
    - Narrative maybe more clustered around center 3-5

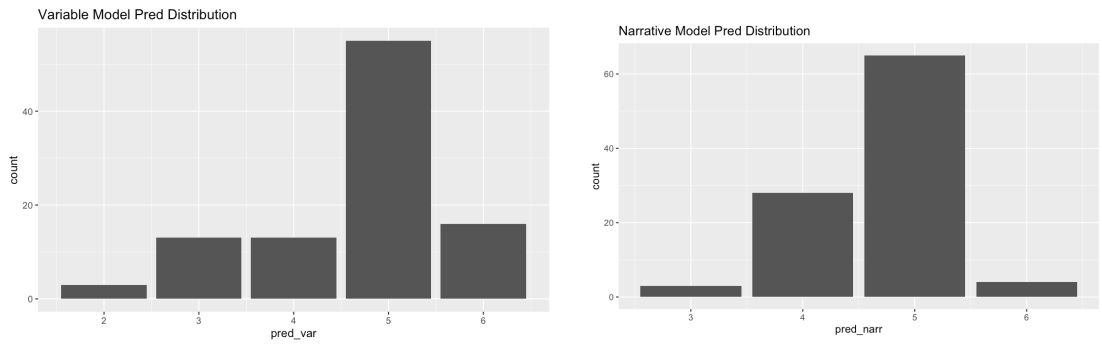
3. Model mislabeled bias

model <chr>	bias <chr>	count <int>	percent <dbl>
Narrative Model	Too Conservative	57	69.51220
Narrative Model	Too Liberal	25	30.48780
Variable Model	Too Conservative	58	68.23529
Variable Model	Too Liberal	27	31.76471

4 rows

- Both models lean conservative on mislabeled cases.





Narrative doesn't have categories 1,2, and 7, clustered towards middle.

- Both models seem to predict more conservatively for all the variables. Particularly age, race and occ10 have large positive mean signed errors (pred - true) > 0.
- Interesting finding: narrative model leans less stereotypical than variable:

- Narrative less skewed to conservative for older age (you would think older is more conservative)
- Narrative more skewed to conservative for black (you would think black is less conservative)
- Narrative less skewed to conservative for more education years (you would think more education is more conservative)
- Narrative more skewed to conservative for single (you would think less conservative for single)
- Narrative less skewed to conservative for South (you would think more conservatives in the South)

Binary Outcome Variable Polviews

I collapse the 7-category POLVIEWS into a binary variable, where 1,2,3,4 are classified as 0 = not conservative and 5,6,7 are classified as 1 = conservative.

Outcome variable: POLVIEWS_BINARY
- Measures individual's political ideology on 2-point discrete scale
- 0 = not conservative; 1 = conservative

Same narrative generation vignettes as before
Duplicated and edited GPT query prompts to make it binary outcome

Results and Discussion:

1. MSE
- Variable MSE 0.53, Narrative MSE 0.61
    - As expected MSE is lower for binary than for seven categories
- Now the Narrative MSE is worse, why?
- Potentially less ambiguity for the variable predictions, narrative models lose the advantage of nuance for 7-categories to make it easier to lean one way or the other
2. F1

Model <chr>	Macro_F1 <dbl>	Weighted_F1 <dbl>
Variable Model	0.4673902	0.4539684
Narrative Model	0.3850187	0.4049440

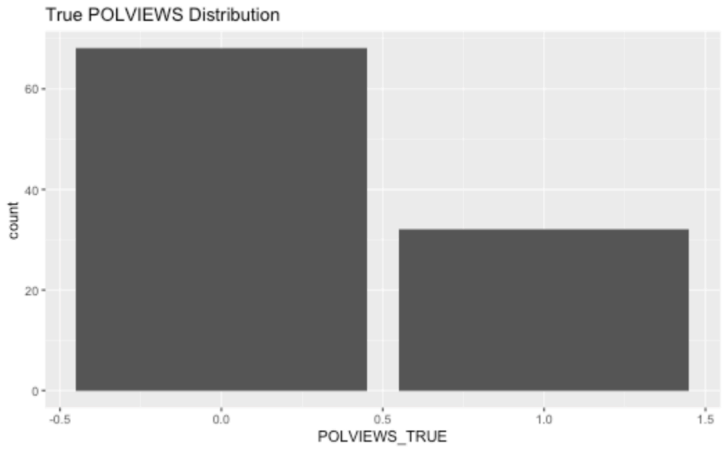
2 rows

- F1 scores are better for variable and worse for narrative once again
3. Model mislabeled bias

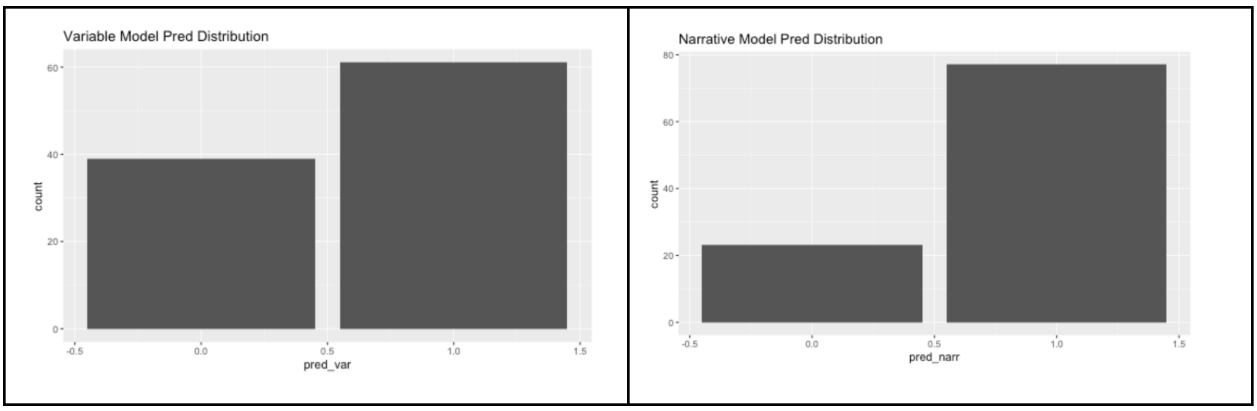
model <chr>	bias <chr>	count <int>	percent <dbl>
Narrative Model	Too Conservative	53	86.88525
Narrative Model	Too Liberal	8	13.11475
Variable Model	Too Conservative	41	77.35849
Variable Model	Too Liberal	12	22.64151

4 rows

- Again both models lean conservative, even more heavily skewed than with 7-categories



-



This is very interesting because the true distribution actually leans not conservative (which makes sense because we group 1,2,3,4 into not conservative and 5,6,7 into conservative so the normal distribution from before had more observations grouped into not conservative), but both models show a heavy skew to conservative, particularly contributed by age, educ, occ10 (errors = 1).

Now looking at the variables,
Narrative is skewed conservatively across all age groups
Narrative skewed more conservative to male now (more stereotypical)
Narrative skewed more conservative to blacks (less stereotypical)
Education has a less interpretable pattern
Narrative skewed more conservative to single (less stereotypical)
Narrative skewed more conservative to South but also more conservative in Midwest (mixed)

Now with binary outcomes, the narrative model is no longer clearly leaning less stereotypical and the results seem mixed. Why?

- Elimination of gradation forces decision making without the degree information, removing the advantage of narrative

- Narrative loses access to those middle ground predictions of 4 and 5 and thus easy to over/undershoot predictions because there are only two categories
- This can perhaps maybe also explain why variable model performance is better on all aspects (MSE and F1)

Why in all cases (binary and 7-categories) do both models lean conservative? possibly:
- Chosen predictors may correlate positively with conservatism in the U.S.
- LLMs are trained on data with stereotypical demographic inference bias that leans towards conservative
- Binary conversion amplifies conservative bias because a single misclassification creates a large error, so if model is slightly conservative-skewed, the error looks really large
- Small sample size

Three Category Outcome Variable Polviews
- Measures individual's political ideology on 3-point discrete scale
- 1 = liberal; 2 = moderate; 3 = conservative

I collapse the 7-category POLVIEWS into a three-category variable, where 1,2,3 are classified as liberal, 4 as moderate, 5,6,7 as conservative.
Outcome variable: POLVIEWS_3

Same narrative generation vignettes as before
Duplicated and edited GPT query prompts to make it binary outcome

Results and Discussion:

1. MSE
- Variable MSE 1.03, Narrative 0.66
- Why does the narrative MSE perform so much better? Likely because it clusters towards middle, reducing error even if the score is not the correct true
2. F1

Model <chr>	Macro_F1 <dbl>	Weighted_F1 <dbl>
Variable Model	0.6888726	0.6822843
Narrative Model	0.6780381	0.6540428

2 rows

- F1 scores are slightly better for variable and worse for narrative (same as binary and 7-category)

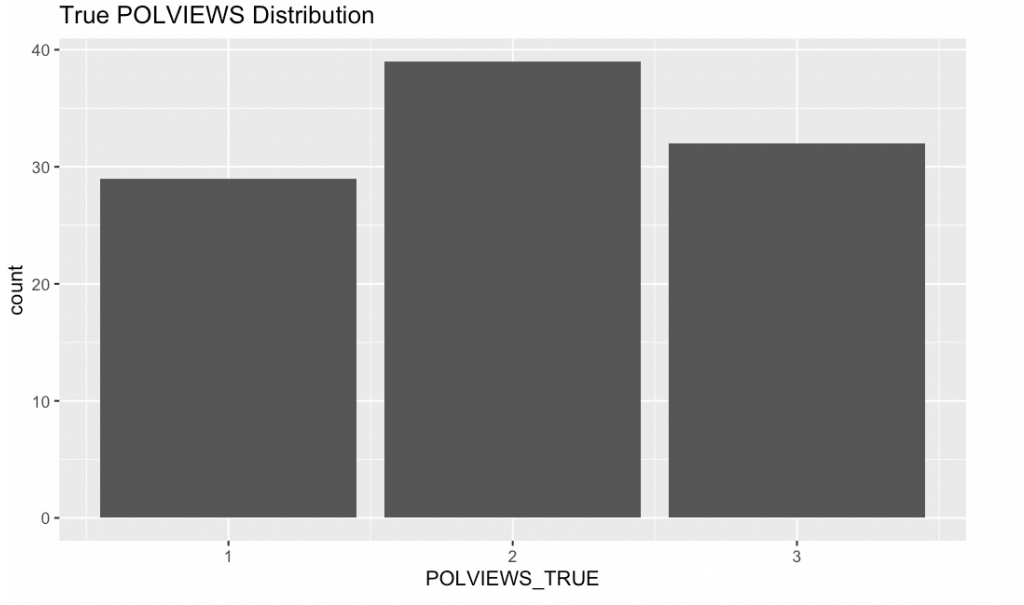
3. Model mislabeled bias

model <chr>	bias <chr>	count <int>	percent <dbl>
Narrative Model	Too Conservative	33	61.11111
Narrative Model	Too Liberal	21	38.88889
Variable Model	Too Conservative	37	60.65574
Variable Model	Too Liberal	24	39.34426

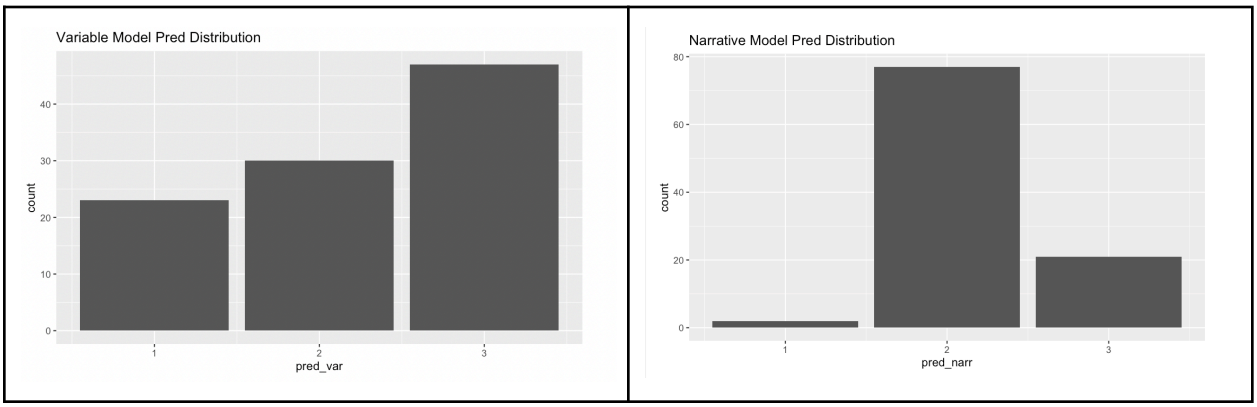
4 rows

- Both models lean conservative as expected (same as binary and 7-category)

The prediction distributions are really interesting here:



True distribution is somewhat balanced/normal, more moderates



- Variable clearly leaning more conservative on 3 while narrative is much more biased towards 2

Now looking at the variables,
Narrative less stereotypical for age (less conservative older people)
Narrative more stereotypical for sex (more conservative males)
Narrative less stereotypical for race (more conservative blacks)

Narrative less stereotypical for educ (less conservative for more educated)
Narrative less stereotypical for marital (more conservative for never married)
Narrative less conservative for most job types
Narrative less stereotypical for region (less conservative in South, Midwest)

→ very similar results to 7-category

Four Category Outcome Variable Polviews

1 = extremely liberal; 2 = slightly liberal; 3 = moderate; 4 = conservative

I collapse the 7-category POLVIEWS into a four-category variable, where 1,2 are classified as extremely liberal, 3 as slightly liberal, 4 as moderate, and 5,6,7 as conservative.
Outcome: POLVIEWS_4

Same narrative generation vignettes as before
Duplicated and edited GPT query prompts to make it binary outcome

Results and Discussion:

1. MSE
- Variable MSE 1.71, Narrative 1.47
- Why does the narrative MSE perform better? Same as above for 3-category, narrative clustering towards middle
2. F1

Model <chr>	Macro_F1 <dbl>	Weighted_F1 <dbl>
Variable Model	0.7652263	0.7261977
Narrative Model	0.7779194	0.7108733

2 rows

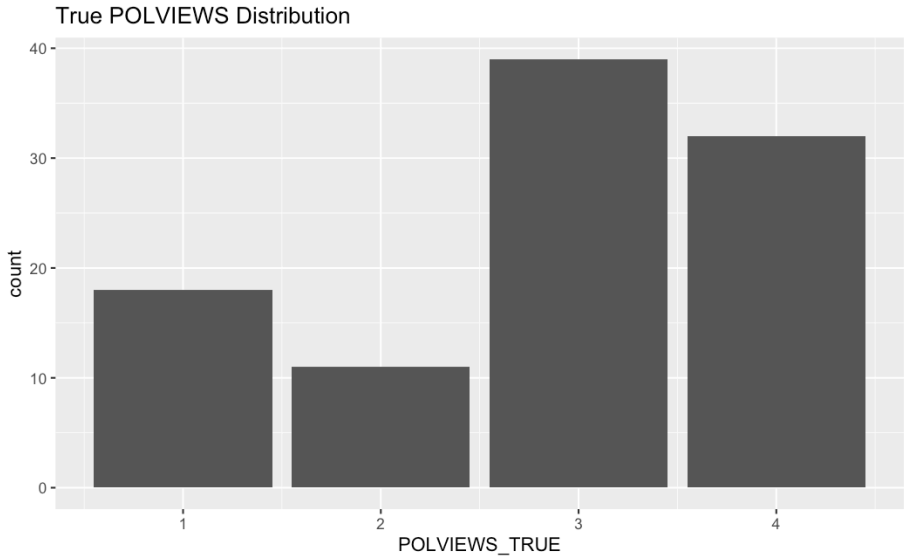
- F1 score is comparable between the two

3. Model mislabeled bias

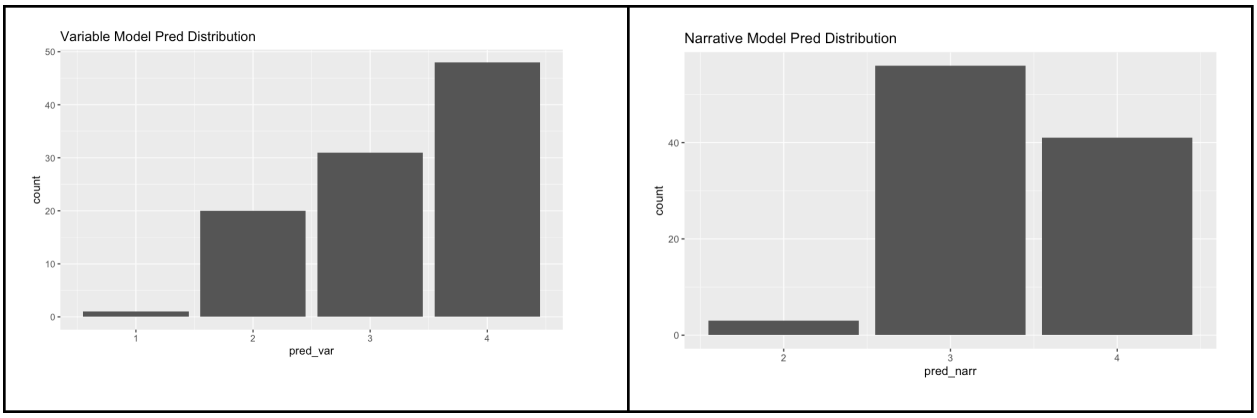
model <chr>	bias <chr>	count <int>	percent <dbl>
Narrative Model	Too Conservative	37	68.51852
Narrative Model	Too Liberal	17	31.48148
Variable Model	Too Conservative	43	66.15385
Variable Model	Too Liberal	22	33.84615

4 rows

Both models still leaning conservative.



3 is moderate, so true distribution somewhat normal



Variable is more spread out than narrative and leans more conservative. Narrative is super clustered towards the moderate and also leans conservative.

Now looking at the variables,
Narrative less stereotypical for age (less conservative older people)
Narrative mixed for sex (more conservative males but also more conservative females)
Narrative less stereotypical for race (more conservative blacks and less conservative whites)
Narrative less stereotypical for educ (less conservative for more educated)
Narrative less stereotypical for marital (more conservative for never married, less conservative for married)
Narrative less conservative for most job types
Narrative mixed stereotypical for region (more conservative in South and midwest, but also more conservative in west)
→ Results are similar to 3-category and 7-category, particularly the patterns seen in plots

Five Category Outcome Variable Polviews

1 = extremely liberal; 2 = liberal; 3 = moderate; 4 = conservative; 5 = extremely conservative

I collapse the 7-category POLVIEWS into a five-category variable, where 1 is classified as extremely liberal, 2,3 as liberal, 4 as moderate, and 5,6 as conservative, 7 as extremely conservative.
Outcome: POLVIEWS_5

Same narrative generation vignettes as before
Duplicated and edited GPT query prompts to make it binary outcome

Results and Discussion:

1. MSE
- Variable MSE 1.43, Narrative MSE 1.23
- Narrative MSE still performs better (same as 4-category, 3-category, 7-category)

2. F1

Model <chr>	Macro_F1 <dbl>	Weighted_F1 <dbl>
Variable Model	0.8051819	0.7118710
Narrative Model	0.8096829	0.7134728

2 rows

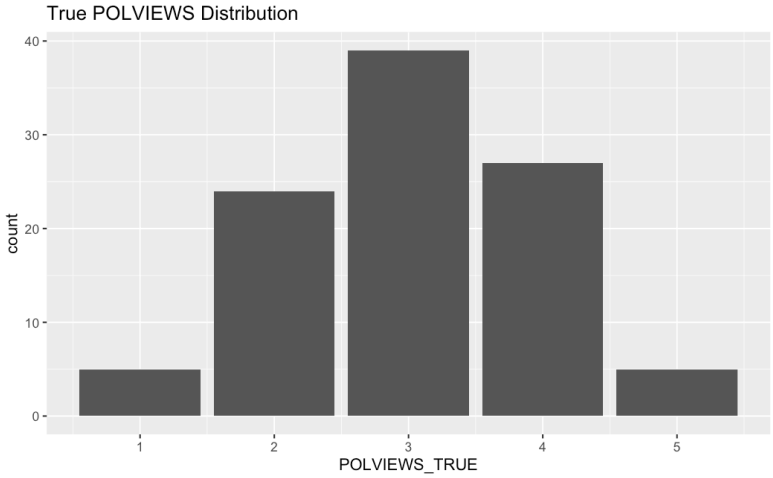
- F1 scores comparable

3. Model mislabeled bias

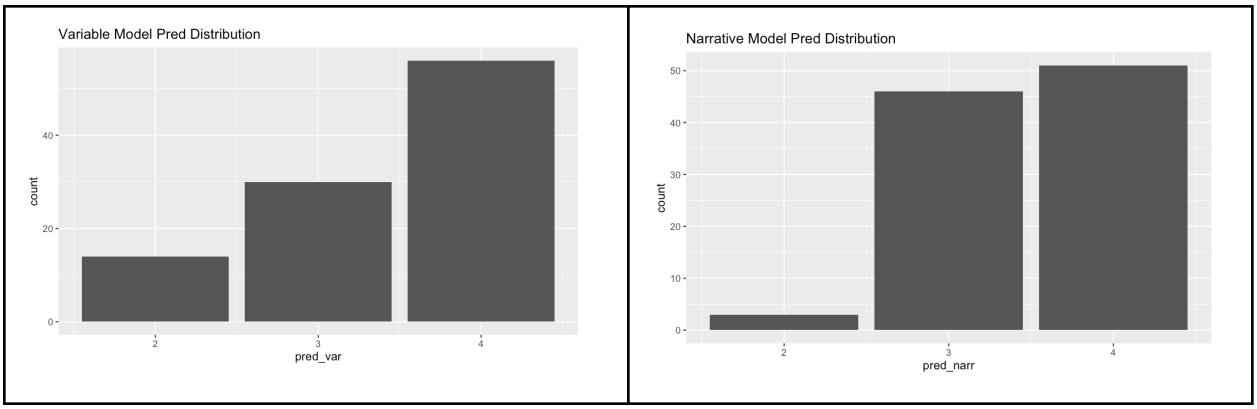
model <chr>	bias <chr>	count <int>	percent <dbl>
Narrative Model	Too Conservative	45	72.58065
Narrative Model	Too Liberal	17	27.41935
Variable Model	Too Conservative	47	72.30769
Variable Model	Too Liberal	18	27.69231

4 rows

Both models lean conservative.



True distribution normal.



Both models do not like the extremes, but the narrative model is still more clustered to 3= moderate than variable.

Now looking at the variables,
Narrative more stereotypical for age (more conservative older people)
Narrative more stereotypical for sex (more conservative males)
Narrative less stereotypical for race (more conservative blacks and less conservative whites)
Narrative less stereotypical for educ (less conservative for more educated)
Narrative less stereotypical for marital (more conservative for never married, less conservative for married)
Narrative less conservative for most job types
Narrative mixed stereotypical for region (less conservative in NE, more conservative in west)

Narrative is somewhat more stereotypical overall for 5-category than for 3 or 4-category
→ Results are still overall similar to the rest of the categories (with the exception of binary)

Summary: Binary seems to be a special case where the narrative model is not as good at prediction than the variable. The other categories (3,4,5,7) have similar results, the narrative has lower MSE yet comparable or worse F1 scores. This could be explained by the narrative clustering towards the middle when >2 categories, but when there are only 2 categories, the narrative loses the gradation advantage. In other words, the gradation advantage is the benefit the narrative model gets from being able to predict middle categories on an ordered scale, allowing for more flexibility (maybe also allows the narrative to better express mixed political signals). Middle predictions reduce error when the model is uncertain. In a binary setting, the middle disappears, so the narrative no longer has this advantage and performs worse. Generally, the narrative is also less stereotypical than the variable for all the categories with the exception of binary where the results are a little more mixed. A model that is less stereotypical is one that hedges its bets (more predictions near the middle as we saw). This explains its good MSE but poorer F1 score in the multi-category settings.

Why does the narrative tend to predict towards the middle but less so for variable?

- Ambiguous signals in narrative
- Narrative has more latent information

3 variable selection

I regenerate the narratives using only three predictors on Chat GPT 5.2 Auto thinking. Unfortunately, when the new GPT model was rolled out on December 11, it superseded the old 5.1 Auto-thinking I used before to generate narratives. I could rerun the old code on new narratives (may take some time to redo). I don't think there would be a major difference between the two models to generate narratives though.

It is interesting to note that the new narratives are much more generic than before and extremely similar across observations, as expected with only three predictors.

Link to narrative generation:
https://chatgpt.com/share/6940c66e-e7a4-800d-862d-3e82e756dafd

Results and Discussion (default 7-category outcome variable):

1. MSE
- Variable MSE 4.28, Narrative MSE 2.47
- Narrative has lower MSE; same as in 7-variable prediction
2. F1 comparable to narrative for macro, better for weighted, same as in 7-var

A tibble: 2 × 3

Model <chr>	Macro_F1 <dbl>	Weighted_F1 <dbl>
Variable Model	0.8378116	0.7558277
Narrative Model	0.8467400	0.6955039

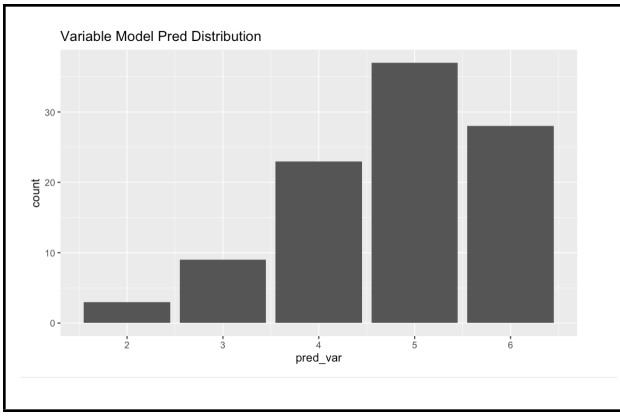
2 rows

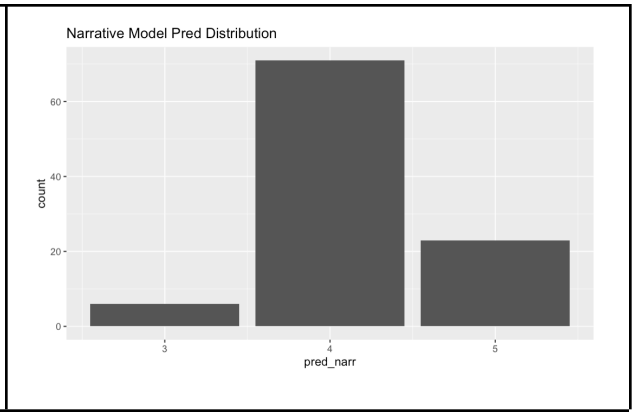
3. Model mislabeled bias; both models still lean conservative but less drastically than in 7-var

model <chr>	bias <chr>	count <int>	percent <dbl>
Narrative Model	Too Conservative	39	57.35294
Narrative Model	Too Liberal	29	42.64706
Variable Model	Too Conservative	60	68.18182
Variable Model	Too Liberal	28	31.81818

4 rows

Narrative is even more drastically centered towards the middle prediction than 7-var





Now looking at the variables,
Narrative generally less conservative for everything

Binary Outcome Variable Polviews

Results and Discussion:

1. MSE
- Variable 0.6, Narrative 0.58
2. F1, comparable in macro, variable better in weighted

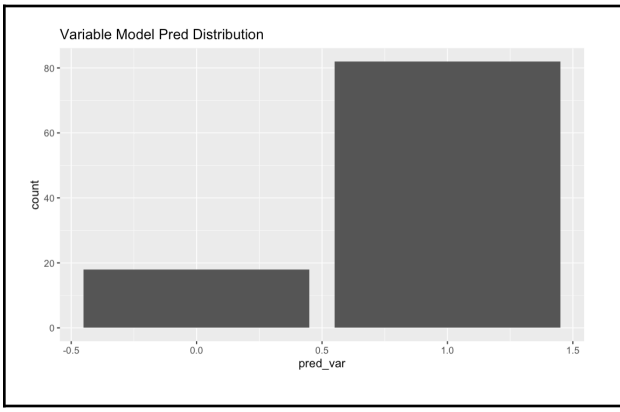
Model <chr>	Macro_F1 <dbl>	Weighted_F1 <dbl>
Variable Model	0.3939394	0.4206061
Narrative Model	0.4047619	0.3628571

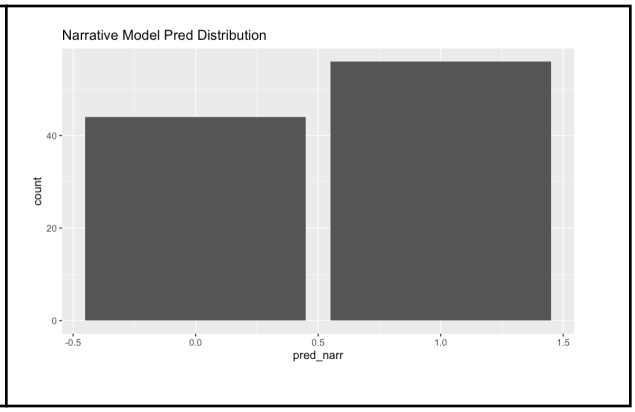
2 rows

3. Model mislabeled bias, conservative leaning for both

model <chr>	bias <chr>	count <int>	percent <dbl>
Narrative Model	Too Conservative	43	74.13793
Narrative Model	Too Liberal	15	25.86207
Variable Model	Too Conservative	57	95.00000
Variable Model	Too Liberal	3	5.00000

4 rows





Narrative is actually more evenly split in this case compared to the binary case in 7-category

Looking at variables,
Narrative less stereotypical for age (more liberal for older)
Narrative mixed stereotypical for race (less conservative for both)
Narrative less stereotypical for race (more conservative for black)

Three Category Outcome Variable Polviews

Results and Discussion:

1. MSE
- Variable 1.5, Narrative 0.63 with much better narrative MSE (same results as 7-var)
2. F1; Variable has higher F1 scores across the board (same as 7-var)

Model <chr>	Macro_F1 <dbl>	Weighted_F1 <dbl>
Variable Model	0.6441004	0.6600717
Narrative Model	0.5671629	0.5091678

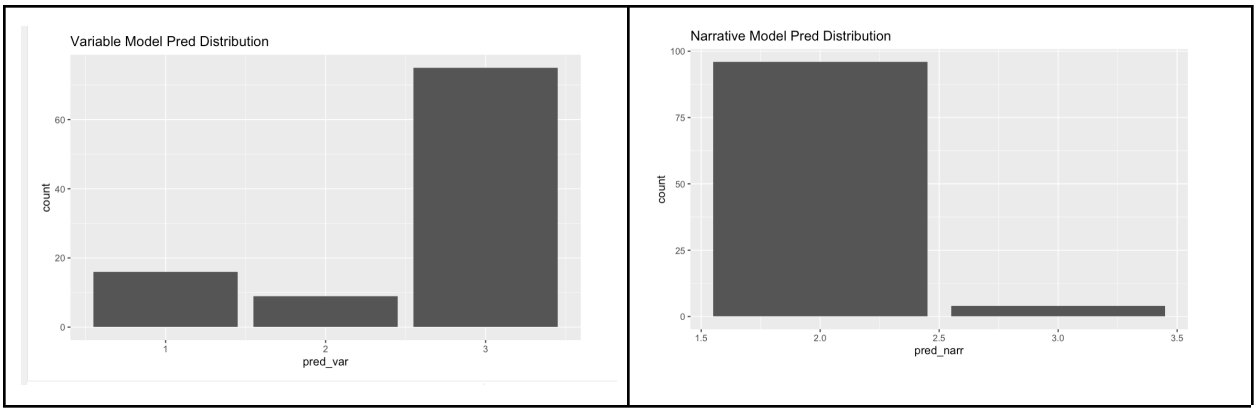
2 rows

3. Model mislabeled bias

model <chr>	bias <chr>	count <int>	percent <dbl>
Narrative Model	Too Conservative	33	55.00000
Narrative Model	Too Liberal	27	45.00000
Variable Model	Too Conservative	52	78.78788
Variable Model	Too Liberal	14	21.21212

4 rows

Both models conservative leaning but variable more drastically skewed conservative. Interestingly, the narrative model predicts no 1s (liberal) and nearly all predictions are 2s (moderate). This is even more moderate-leaning behavior than in 7-var, likely because there is practically no latent information for the model to interpret from only age, sex and race. When you look at the vignettes, that makes sense because the vignettes are very homogenous across all individuals. So the narrative has a better MSE only because it makes centered predictions, not because it accurately predicts the correct answer.



Looking at variables, narrative is less conservative across the board. I think generally because the narrative model pretty much only predicts 2s moderates, the mean signed error is smaller than the variable since the variable makes more varied predictions.

Four Category Outcome Variable Polviews

Results and Discussion:

1. MSE

- Variable 2.38, Narrative 1.19 (same as 7-var and other categories except binary)

2. F1

Model <chr>	Macro_F1 <dbl>	Weighted_F1 <dbl>
Variable Model	0.7329950	0.7007251
Narrative Model	0.6874646	0.5666344

2 rows

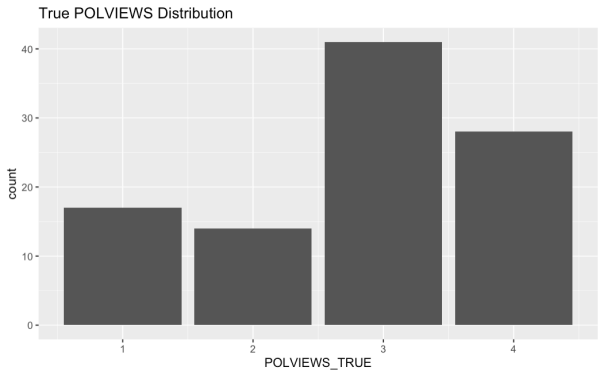
Variable significantly better F1 scores, same as 7-var

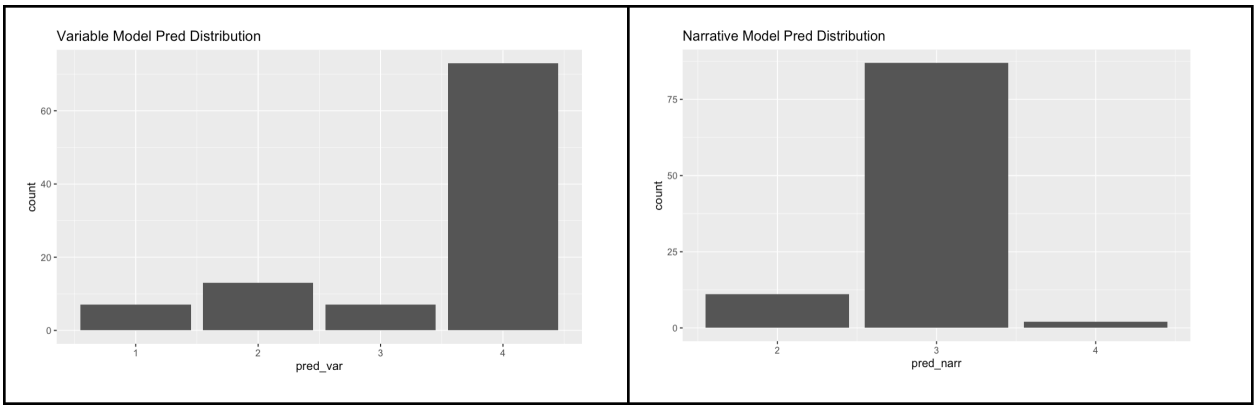
3. Model mislabeled bias

model <chr>	bias <chr>	count <int>	percent <dbl>
Narrative Model	Too Conservative	31	47.69231
Narrative Model	Too Liberal	34	52.30769
Variable Model	Too Conservative	51	72.85714
Variable Model	Too Liberal	19	27.14286

4 rows

Now this is pretty interesting. The narrative model is now skewed liberal, although by not much. The narrative model pretty much predicts everyone at moderate so that there are just few conservatives and liberals in general. One potential explanation is that with less latent information the narrative model relies more heavily on its prior to build its posterior. In other words, narratives flatten likelihoods because many lifestyles are observationally equivalent





Variable specific bias shows similar results as three-category.

Five Category Outcome Variable Polviews

Results and Discussion:

1. MSE

- Variable 1.75, Narrative 1.12, Narrative is better (same as 7-var)

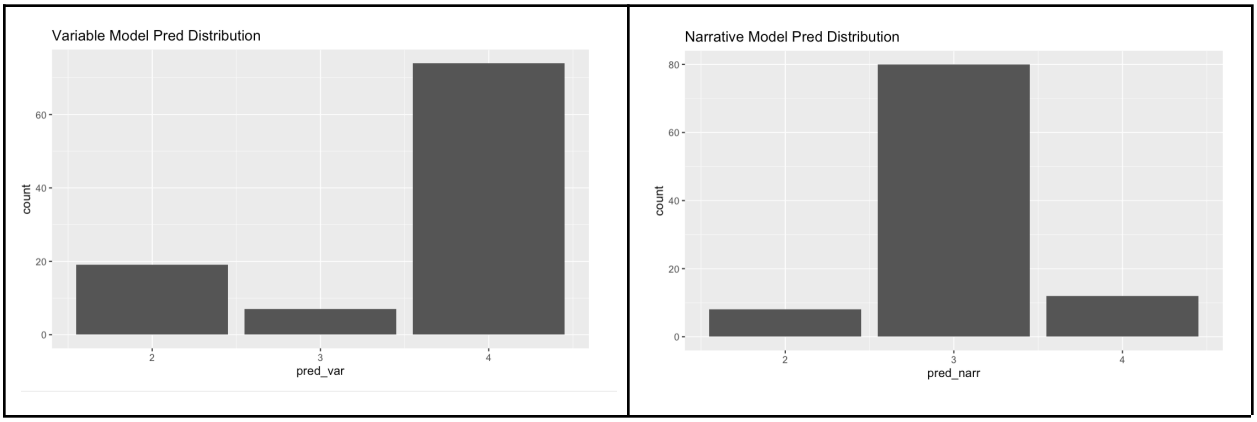
2. F1; Variable is better

Model <chr>	Macro_F1 <dbl>	Weighted_F1 <dbl>
Variable Model	0.7820460	0.7024562
Narrative Model	0.7612083	0.5868951

2 rows

3. Model mislabeled bias; both models lean conservative

model <chr>	bias <chr>	count <int>	percent <dbl>
Narrative Model	Too Conservative	35	53.03030
Narrative Model	Too Liberal	31	46.96970
Variable Model	Too Conservative	52	73.23944
Variable Model	Too Liberal	19	26.76056



For specific variables,

Narrative less stereotypical for age (less conservative for older but more conservative for younger)
Narrative less stereotypical for sex (fewer male conservatives than female)
Narrative less stereotypical for race (more conservative blacks)

Summary: 3-variable results are generally similar to 7-variable. First, the narrative model as a whole still makes moderate-leaning predictions compared to the variable. Second, both models generally still lean conservative. One difference is that the 3-variable results have less spread than 7-variable results, particularly in the narrative model. A larger percentage of observations in the narrative model are predicted as moderate. Another surprising finding is that in the 4-category outcome variable, the narrative model actually was leaning slightly liberal, but it was leaning conservative for everything else. In the 4-category outcome, there is a slight unevenness in the true distribution because 1,2 are both liberal, 3 is moderate and 4 is conservative. This could potentially lead to predictions that are also slightly more liberal-biased in the 4-category scheme only. Another potential explanation to keep in mind for the narrative model is that having fewer predictors means less latent information for the narrative advantage and more homogeneity in the narrative predictions which may explain the behavior seen.

Literature reviews

Are there existing similar publications exploring variable vs narrative model?

- Not exactly the same, as far I can tell

But there are adjacent publications:

This one is maybe most similar and perhaps relevant to what we feed into the prompt: "Towards Better Serialization of Tabular Data for Few-Shot Classification with Large Language Models" (Jaitly, et al., 2023)

Key results: They tried 4 different ways to convert tabular data into text before feeding it to an LLM (Chat 3.5 and 4).

1. **Plain Natural-Language Serialization**
   Example:
   "A 45-year-old male with income 50000 and education 16 years…"

2. **More Structured Serialization**
   Example:
   "age: 45; gender: male; income: 50000; education: 16; …"

3. **LaTeX Table Serialization**
   Encoding the row into a LaTeX-style table snippet
   (surprisingly this helps because the structure is explicit).

4. **Attribute Sorting / Reordering**
   Reordering variables alphabetically or semantically to mimic training distributions.

LateX performed the best, followed by structured then natural language.

Other literature:

1. Variable model prediction

- TabLLM: Few-shot Classification of Tabular Data with Large Language Models (Hegselmann et al., 2023)
- Key results: They show that even with no additional training, LLM gives non-trivial performance on tabular classification. It shows that LLMs have general pretrained "world knowledge" that can transfer to structured-data tasks.

2. Narrative model prediction

- Can large language models help predict results from a complex behavioural science study? (Lippert et al., 2024)
- Key results: They used GPT-3.5 and GPT-4. In Study 1, they gave the LLMs the design of a large-scale experiment (about emotions, gender, social perception) and asked them to predict empirical effect sizes. They compared to 119 human experts. They found GPT-4 matched the human experts (correlation ~0.89) while GPT-3.5 did poorly (~0.07). In Study 2, participants could query a GPT-4 powered chatbot and that improved their forecasting accuracy. Essentially forecasting behavioral science results that somewhat mimic narrative

Do LLMs demonstrate chain of thought?

Papers of support:

Language models, like humans, show content effects on reasoning tasks (Lampinen et. al 2024)

The authors compared the reasoning performance of humans and LLMs on **three types** of logic/reasoning tasks:

1. Natural Language Inference (NLI)

2. Syllogism-validity judgments

3. The Wason selection task

They evaluated multiple LLMs and compared to humans, especially looking at how much semantic content ("content effect") influenced accuracy and other patterns (like confidence, response-time for humans, et

Key result: Both humans and LLMs show strong content effects, accuracy is higher when the semantic content supports the correct logical inference than when content conflicts or is

arbitrary.
In the syllogism and Wason tasks in particular, both humans & models perform better when the content aligns with prior real-world knowledge.

In some cases, LLMs performed *better than humans* (especially in harder tasks like Wason) and exhibited similar but not identical patterns of error.

They also found that LLM confidence correlates in similar ways to human response-time patterns: when humans take longer to respond, accuracy is lower. Similarly LLMs express lower confidence in cases where humans take time.

LLMs don't just apply "pure logic" independently of meaning, they inherit human-like biases in reasoning through semantic content.

LLMs used: GPT-4, GPT-3.5, PaLM-2, LLaMA-2, Claude
Does not directly address Turpin critical paper that came before this

Chain-of-Though Prompting Elicits Reasoning in Large Language Models (Wei et al. 2022)

Key results: first demonstration that reasoning can be elicited through prompting alone. They use chain of thought prompting, which is showing the model worked examples . When you *demonstrate* step-by-step reasoning in the prompt, large models begin generating their own reasoning chains. Dramatically boosts accuracy (e.g., >40% absolute improvement on difficult math).Small models (<10B) don't benefit much which suggests reasoning emerges at scale.

LLMs used: five models - original GPT-3, LaMDA, PaLM, UL2 20B, Codex

Large Language Models are Zero Shot Reasoners (Kojima et. al 2022)

Key results: LLMs already internally have encoding procedures for chain of thought. Zero-shot CoT yields performance similar to multi-example CoT. Add phrase "lets think step by step"

LLMs used: 17 models total, main ones are Instruct GPT-3, original GPT-3, GPT-2, OPT-175B, BLOOM, PaLM
Released right after Wei et. al, as a comlement

Tree of Thoughts: Deliberate Problem Solving with Large Language Models (Yao et al. 2023)

Key results: Extend CoT from one linear chain to a branching search tree. The model explores several partial thoughts, evaluates them, backtracks, and expands promising ones. Greatly improves performance on planning-heavy tasks (e.g., puzzles like "Game of 24"). Better than CoT when the problem requires trying multiple strategies. Key is that ToT helps backtrack thinking."

LLMs used: main is GPT-4, some GPT-3.5

Papers of criticism:
"Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting" Turpin et al. (2023)

Key results: They give the model hard reasoning tasks. The final answer does not match CoT reasoning. Challenges the interpretation of CoT as "how the model really thinks. Findings are that CoT is often a post-hoc rationalization rather than actual internal reasoning.
For ex, If you bias the input, the model's explanation often omits the influence of that bias even when its answer reflects the bias

LLMs used: GPT-3.5, claude-v1.0
Directly criticizes Wei et al. (2022) and Kojima et al. (2022).
A leading critique paper

Preemptive Answer "Attacks" on Chain-of-Thought Reasoning (Xu et al., 2024)

Key results: They introduce a scenario they call *"preemptive answers"*: the model receives or infers the correct answer *before* it attempts the chain of thought reasoning.
They show that when a model obtains this preemptive answer, its subsequent reasoning is significantly worse. The chain of thought may simply rationalize the answer rather than actually reason through it.

LLMs used: GPT-4
New and gaining traction, extends Turpin et al. (2023).

Which model is best at CoT reasoning?

Interesting article →
https://composio.dev/blog/cot-reasoning-models-which-one-reigns-supreme-in-2025

On ChatGPT o1:

"Similar to how a human may think for a long time before responding to a difficult question, o1 uses a chain of thought when attempting to solve a problem. Through reinforcement learning, o1 learns to hone its chain of thought and refine the strategies it uses. It learns to recognize and correct its mistakes. It learns to break down tricky steps into simpler ones. It learns to try a different approach when the current one isn't working. This process dramatically improves the model's ability to reason"
"model appears iteratively to evaluate each step and to redirect its own search in the solution space for the problem."

https://blog.iese.edu/artificial-intelligence-management/2024/chain-of-thought-reasoning-the-new-llm-breakthrough/

How does CoT work?
Definition: the model uses intermediate reasoning steps to get a final answer

- Train model

Wei et. al. (2022) show increasing model scale improves CoT due to a mixture of various reasons including semantic understanding, stronger symbol mapping, stronger arithmetic abilities, faithfulness, etc. They find that large GPT and PalM models perform the best with CoT prompting for arithmetic, common sense and symbolic reasoning tasks. CoT prompting means instead of fine tuning the model itself, one can prompt a few examples aka “few shot prompting”.

What makes a model good at CoT?

1. RL rewarding good thinking/deliberation
2. Reward function gives high values to the process “Process supervision”
3. “Test-time compute”: multiple solution paths, longer thinking time.. Extra processing power during inference
    a. This can be scaled like o1 does: o1 dynamically increases their reasoning time during inference. This means they can spend more time thinking about complex questions, improving accuracy at the cost of higher compute usage.

DeepSeek transferred reasoning skills from their R1 model to smaller AI models, called distillation, such that both large and small open-source models show good performance. (https://huggingface.co/blog/Kseniase/testtimecompute)

Most recent model introductions and updates (as of Dec 17, 2025):

Claude 4 [Anthropic] (May 2025)

- Claude Opus 4 for coding
- Overall best for agentic coding
- Not as good at Olympiad math

GPT 5 [OpenAI] (Aug 2025); GPT 5.1 [OpenAI] (Nov 2025); GPT 5.2 [OpenAI] (Dec 2025)

- 5.2 builds on long context reasoning, it has a 5.2 ‘Thinking’ version
- Best for professional work and multi-step projects

Gemini 3 [Deepmind] (Nov 2025); Gemini Flash [Deepmind] (Dec 2025)

- Also Gemini 3 Deep Think version for deeper reasoning on harder tasks
- Flash for reasoning but faster

Llama 4 [Meta] (Apr 2025)

- Not very good and not mainstream

DeepSeek R1 (reasoning focused) (Jan 2025); DeepSeek-V3.2 (Dec 2025)

- R1 is explicitly trained to incentivize reasoning behaviors

Why do we see observed patterns in narrative vs variable? Does CoT help explain?

Hypothesis: CoT may help amplify narratives to lead to better narrative results when the narratives provide richer inputs and have more latent information for CoT to use.