Can ChatGPT (and Friends) Specify and Estimate a Multinomial Logit Model?

Georges Sfeir, Gabriel Nova

Generating insights in socio-technical systems using LLMs

TU Delft - CityAI Lab

June 23, 2025

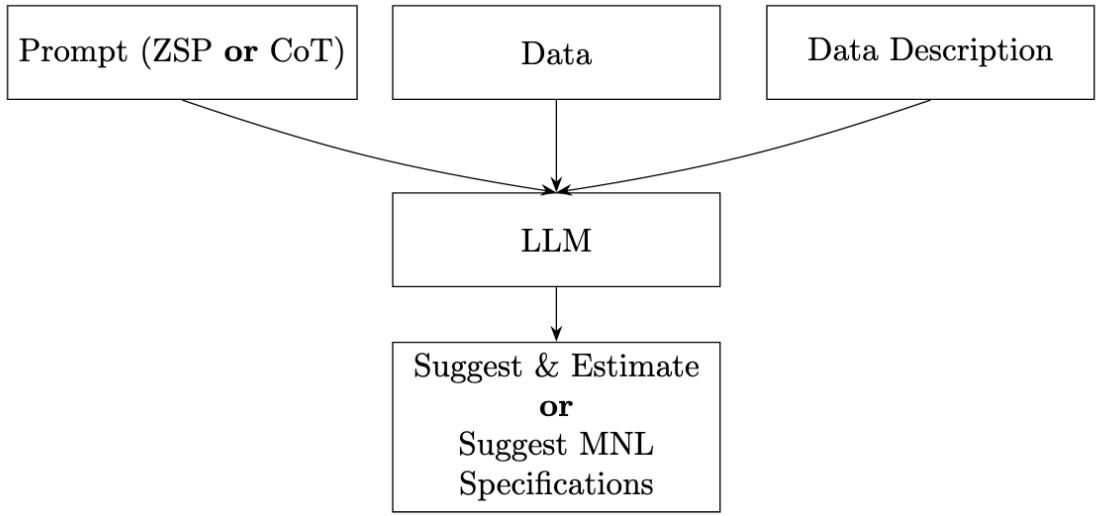
Introduction / Motivation

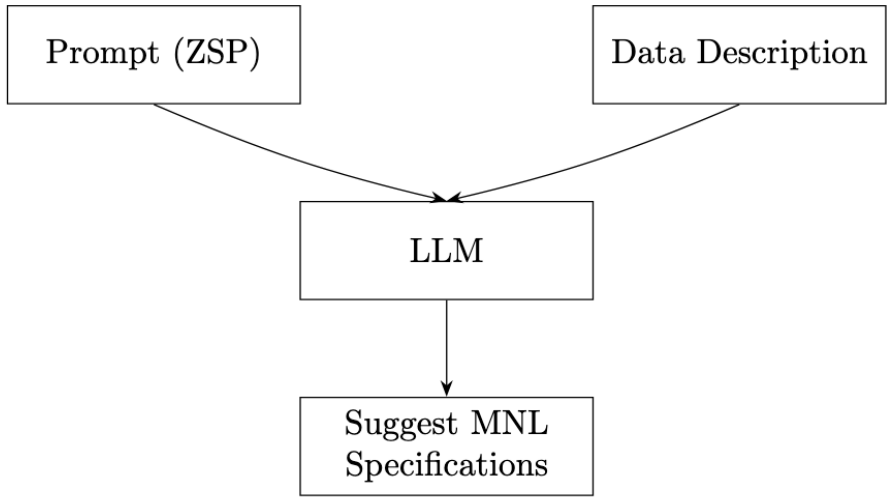
- Multinomial Logit (MNL) Models are a core tool in discrete choice analysis
- Specifying MNLs requires expertise, domain knowledge, and is time consuming:
  - Which variables to include?
  - What transformations to apply?
  - Should we include interactions terms between attributes and covariates etc.?
- Meanwhile, Large Language Models (LLMs) are increasingly capable of:
  - Understanding structure data
  - Generating code
  - Emulating expert reasoning

Introduction / Motivation

- Can LLMs correctly specify and estimate MNL models?
- This study evaluates the ability of 9 LLMs to automatically specify and estimate MNLs across different:
  - Prompting strategies
    - Zero-Shot
    - Chain-of-Thoughts
  - Information settings
    - Data & Data Description
    - Data Description
  - Modelling goals
    - Suggest & Estimate Specifications
    - Suggest Specifications

Experimental Framework





Experiment	Information Setting	Prompting Strategy	Modelling Goal
1	Data & Data Description	ZSP	Suggest & Estimate
2	Data & Data Description	CoT	Suggest & Estimate
3	Data & Data Description	ZSP	Suggest
4	Data & Data Description	CoT	Suggest
5	Data Description	ZSP	Suggest

ZSP: Zero-Shot Prompt

CoT: Chain-of-Thoughts

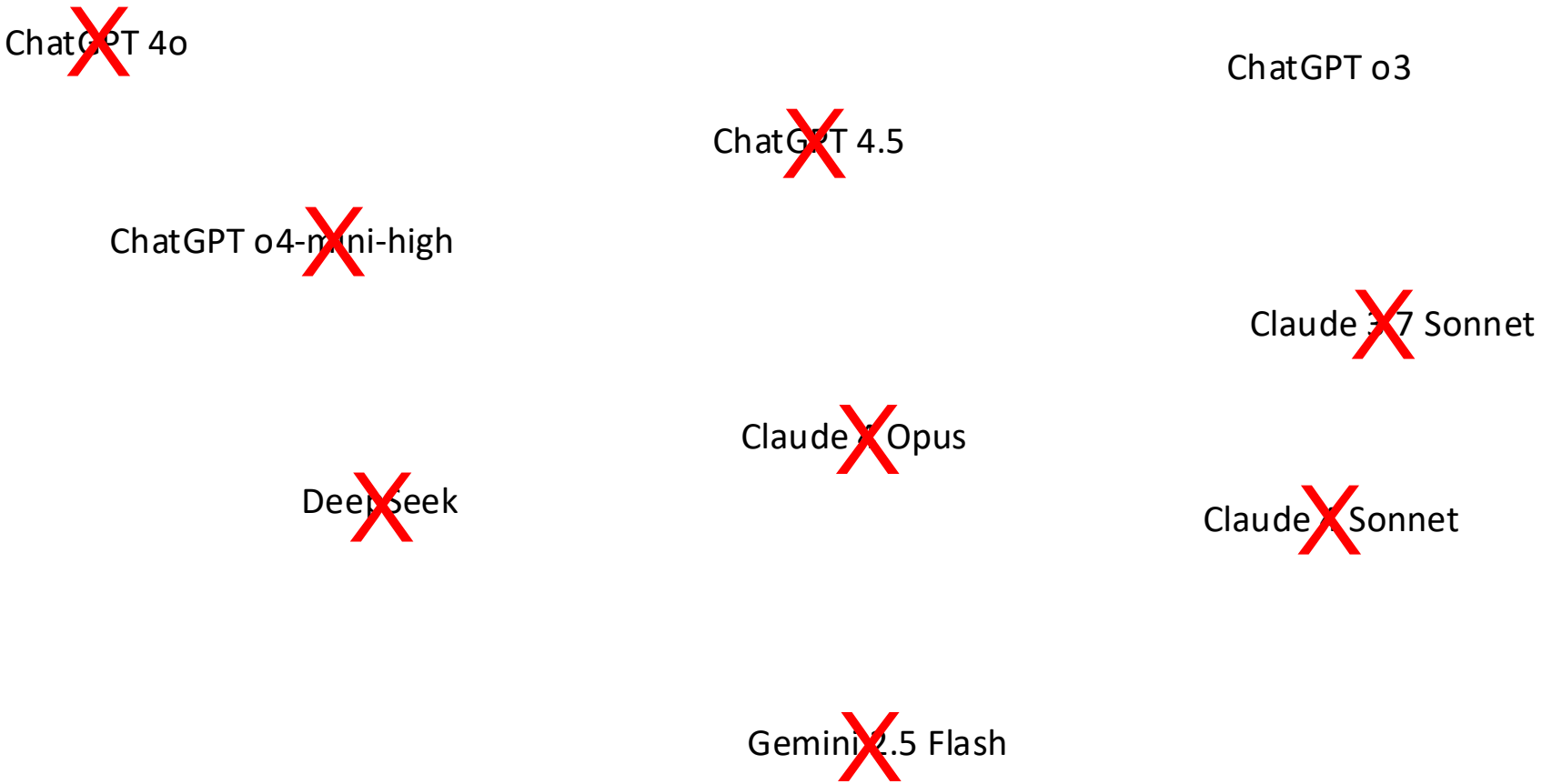
Application

- Dataset
  - Apollo Mode Choice
  - Number of Individuals: 500
  - Number of Observations: 1,000
  - Alternatives: Car, Bus, Air, Rail
- LLMs:
  - ChatGPT 4o, o4-mini-high, o3, 4.5
  - Claude 4 Opus, 4 Sonnet, 3.7 Sonnet
  - DeepSeek
  - Gemini 2.5 Flash
- Total number of specifications generated: 152

Experiments 1 & 2

Experiment	Information Setting	Prompting Strategy	Goal
1	Data & Data Description	ZSP	Suggest & Estimate
2	Data & Data Description	CoT	Suggest & Estimate
3	Data & Data Description	ZSP	Suggest
4	Data & Data Description	CoT	Suggest
5	Data Description	ZSP	Suggest

Results – Experiments 1&2: Suggest & Estimate



choice modelling centre

Results – Experiments 1&2: Suggest & Estimate

Experiment 1 (Full Information, Zero-Shot, Suggest & Estimate)

Specification (ChatGPT o3)	LL	AIC	BIC
S1	−1,031.84	2,073.68	2,098.21
S2	**−981.80**	**1,977.61**	**2,011.96**
S3	−1,083.46	2,176.91	2,201.45

Experiment 2 (Full Information, Chain-of-Thoughts, Suggest & Estimate)

Specification (ChatGPT o3)	LL	AIC	BIC
S1	−1,031.82	2,073.63	2,098.17
S2	−1,030.97	2,073.93	2,103.38
S3	−1,026.81	2,065.62	2,095.06
S4	**−981.80**	**1,977.61**	**2,011.96**
S5	−993.30	2,000.59	2,034.95

choice modelling centre

Experiment 3

Experiment	Information Setting	Prompting Strategy	Goal
1	Data & Data Description	ZSP	Suggest & Estimate
2	Data & Data Description	CoT	Suggest & Estimate
3	Data & Data Description	ZSP	Suggest
4	Data & Data Description	CoT	Suggest
5	Data Description	ZSP	Suggest

Results – Experiment 3: Data & Data Description, ZSP, Suggest

Log-Likelihood

Specification	ChatGPT 4o	ChatGPT o4-mini-high	ChatGPT o3	ChatGPT 4.5	Claude 4 Opus	Claude 4 Sonnet	Claude 3.7 Sonnet	DeepSeek R1	Gemini 2.5 Flash
S1	−1,106.23*	−1,031.82	−1,031.00	−1,030.97	−1,030.97	−1,031.00	−1,030.98	−1,089.94*	−1,031.82
S2	−1,027.57*	−1,030.97	−1,030.97	−1,024.48	−1,048.30†	−1,025.91	−1,025.00	−1,025.82*	−1,030.97
S3	−1,022.32*	−1,025.00	−1,036.49	**−974.95**	−1,026.10	−1,017.58	**−978.15**	−1,080.08*	**−983.45**
S4	−1,023.33*	**−999.72**	−1,011.22	−1,024.18	**−982.15**	**−978.37**	−1,025.36	−68,386.87†	−1,028.78‡
S5	−1,123.26*	−1,036.38	**−977.37**	–	−1,033.34†	**−978.37**	−4,991.18†	−1,026.19*	–
S6	–	–	–	–	−1,027.99	−1,030.78	−967.80‡	−963.94*	–

*No ASCs included. †Model did not converge. ‡Positive Beta Cost and/or Beta Time.

AIC

Specification	ChatGPT 4o	ChatGPT o4-mini-high	ChatGPT o3	ChatGPT 4.5	Claude 4 Opus	Claude 4 Sonnet	Claude 3.7 Sonnet	DeepSeek R1	Gemini 2.5 Flash
S1	2,216.47*	2,073.63	2,071.99	2,073.94	2,073.93	2,071.99	2,073.95	2,185.87*	2,073.63
S2	2,071.14*	2,073.94	2,073.93	2,076.97	2,108.60†	2,065.82	2,074.01	2,073.65*	2,073.93
S3	2,060.64*	2,074.01	2,084.98	**1,987.91**	2,070.19	2,053.15	**1,978.30**	2,166.17*	**1,986.90**
S4	2,062.67*	**2,029.44**	2,036.44	2,082.36	**1,980.31**	**1,976.75**	2,068.72	136,783.74†	2,073.56‡
S5	2,250.53*	2,084.77	**1,984.74**	–	2,090.67†	**1,976.75**	10,016.36†	2,064.38*	–
S6	–	–	–	–	2,067.98	2,073.56	1,963.60‡	1,961.87*	–

*No ASCs included. †Model did not converge. ‡Positive Beta Cost and/or Beta Time.

choice modelling centre

Experiment 4

Experiment	Information Setting	Prompting Strategy	Goal
1	Data & Data Description	ZSP	Suggest & Estimate
2	Data & Data Description	CoT	Suggest & Estimate
3	Data & Data Description	ZSP	Suggest
4	Data & Data Description	CoT	Suggest
5	Data Description	ZSP	Suggest

Results – Experiment 4: Data & Data Description, CoT, Suggest

Log-Likelihood

Specification	ChatGPT 4o	ChatGPT o4-mini-high	ChatGPT o3	ChatGPT 4.5	Claude 4 Opus	Claude 4 Sonnet	Claude 3.7 Sonnet	DeepSeek R1	Gemini 2.5 Flash
S1	−1,024.48	−1,030.97	−1,030.97	−1,031.00	−1,024.48	−1,019.18	−1,024.48	−1,030.97	−1,031.42
S2	**−979.47**	−1,024.48	**−980.62**	−978.34	−1,030.97	−1,010.30	−1,030.97	**−1,025.76**	−1,024.36‡
S3	−1,024.97	**−987.79**	−1,026.00	−1,037.28	**−973.34**	−1,019.76	−1,030.96†	−1,083.58†	**−968.29**
S4	−1,017.16‡	−1,036.49	−1,035.45	−1,030.97	−1,035.14	−1,024.00	**−979.02**	−993.95†	−1,024.48
S5	–	−1,014.87	–	**−975.11**	−1,040.96	**−967.53**	−1,035.94	−1,499.58	−1,030.97
S6	–	–	–	–	−980.82	−1,019.32	−1,021.88	–	−960.07‡
S7	–	–	–	–	–	–	−981.80	–	–

†Model did not converge. ‡Positive Beta Cost and/or Beta Time.

AIC

Specification	ChatGPT 4o	ChatGPT o4-mini-high	ChatGPT o3	ChatGPT 4.5	Claude 4 Opus	Claude 4 Sonnet	Claude 3.7 Sonnet	DeepSeek R1	Gemini 2.5 Flash
S1	2,076.96	2,073.93	2,073.94	2,071.99	2,076.96	2,056.37	2,076.95	2,073.93	2,074.84
S2	**1,992.94**	2,076.95	**1,979.23**	1,970.68	2,073.93	2,038.61	2,073.93	**2,069.51**	2,080.72‡
S3	2,077.94	**1,993.59**	2,070.01	2,084.57	**1,970.69**	2,055.52	2,075.93†	2,179.16†	**1,966.59**
S4	2,068.32‡	2,084.98	2,082.89	2,073.93	2,082.28	2,066.01	1,978.04	2,003.90†	2,076.95
S5	–	2,043.74	–	**1,966.22**	2,091.92	**1,963.05**	2,083.87	3,011.16	2,079.93
S6	–	–	–	–	1,977.63	2,052.64	2,057.76	–	1,972.14‡
S7	–	–	–	–	–	–	**1,977.61**	–	–

†Model did not converge. ‡Positive Beta Cost and/or Beta Time.

choice modelling centre

Experiment 5

Experiment	Information Setting	Prompting Strategy	Goal
1	Data & Data Description	ZSP	Suggest & Estimate
2	Data & Data Description	CoT	Suggest & Estimate
3	Data & Data Description	ZSP	Suggest
4	Data & Data Description	CoT	Suggest
5	Data Description	ZSP	Suggest

Results – Experiment 5: Data Description, ZSP, Suggest

Log-Likelihood

Specification	ChatGPT 4o	ChatGPT o4-mini-high	ChatGPT o3	ChatGPT 4.5	Claude 4 Opus	Claude 4 Sonnet	Claude 3.7 Sonnet	DeepSeek R1	Gemini 2.5 Flash
S1	−1,031.00	−1,121.16*	−1,031.00	−1,031.00	−1,031.82	−1,030.97	−1,030.97	−1,047.71	−1,031.82
S2	**−969.73**	−1,030.97	−1,030.97	−1,030.97	−1,030.97	−1,026.81	−1,024.48	−1,036.03	−1,030.97
S3	−976.60	−1,036.47	−1,020.74	−1,035.77	−1,022.61	−1,023.22	−1,038.05†	**−982.10**	**−981.57**
S4	–	**−981.12**	**−982.49**	**−991.62**	**−972.40**	**−971.33**	**−990.18**	−1,071.10†	−1,024.48
S5	–	−1,018.80	−1,020.41	−994.39	−1,035.14	−978.89	−1,013.76†	−1,025.00	–
S6	–	–	–	–	−1,030.26	−972.07	−1,030.29	−1,046.18	–
S7	–	–	–	–	−1,006.36	–	−1,027.52	–	–
S8	–	–	–	–	–	–	−976.31†	–	–

*No ASCs included. †Model did not converge.

AIC

Specification	ChatGPT 4o	ChatGPT o4-mini-high	ChatGPT o3	ChatGPT 4.5	Claude 4 Opus	Claude 4 Sonnet	Claude 3.7 Sonnet	DeepSeek R1	Gemini 2.5 Flash
S1	2,071.99	2,246.32*	2,071.99	2,071.99	2,073.64	2,073.94	2,073.94	2,107.42	2,073.64
S2	**1,973.46**	2,073.94	2,073.93	2,073.93	2,073.94	2,065.62	2,076.96	2,084.06	2,073.94
S3	1,987.19	2,084.93	2,057.48	2,081.54	2,059.22	2,064.44	2,092.10†	**1,978.20**	**1,983.14**
S4	–	**1,980.24**	**1,990.98**	**2,005.25**	1,960.80	1,972.66	**2,000.36**	2,154.20†	2,076.96
S5	–	2,053.61	2,058.82	2,012.78	2,082.28	1,987.78	2,047.52†	2,074.00	–
S6	–	–	–	–	2,078.52	**1,972.14**	2,078.58	2,106.36	–
S7	–	–	–	–	2,028.72	–	2,067.04	–	–
S8	–	–	–	–	–	–	1,986.62†	–	–

*No ASCs included. †Model did not converge.

choice modelling centre

Best Specifications per Experiment and/or LLM

Results – Best LL Per Experiment Per LLM

Log-Likelihood

	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5
GPT 4o	–	–	–	−979.47	−969.73
GPT o4minihigh	–	–	−999.72	−987.79	−981.12
GPT o3	−981.80	−981.80	−977.37	−980.62	−982.49
GPT 4.5	–	–	−974.95	−975.11	−991.62
Claude 4 Opus	–	–	−982.15	−973.34	−972.40
Claude 4 Sonnet	–	–	−978.37	−967.53	−971.33
Claude 3.7 Sonnet	–	–	−978.15	−979.02	−990.18
DeepSeek	–	–	–	−1,025.76	−982.10
Gemini 2.5 Flash	–	–	−983.45	−968.29	−981.57

Exp. 1: Full/ZS/Estimate, Exp. 2: Full/CoT/Estimate, Exp. 3: Full/ZS/Suggest, Exp. 4: Full/CoT/Suggest, Exp. 5: Limited/ZS/Suggest.

Results – Best AIC Per Experiment Per LLM

	AIC				
	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5
ChatGPT 4o	–	–	–	1,992.94	1,973.46
ChatGPT o4minihigh	–	–	2,029.44	1,993.59	1,980.24
ChatGPT o3	1,977.61	1,977.61	1,984.74	1,979.23	1,990.98
ChatGPT 4.5	–	–	1,987.91	1,966.22	2,005.25
Claude 4 Opus	–	–	1,980.31	1,970.69	1,960.80
Claude 4 Sonnet	–	–	1,976.75	1,963.05	1,972.14
Claude 3.7 Sonnet	–	–	1,978.30	1,977.61	2,000.36
DeepSeek	–	–	–	2,069.51	1,978.20
Gemini 2.5 Flash	–	–	1,986.90	1,966.59	1,983.14

Exp. 1: Full/ZS/Estimate, Exp. 2: Full/CoT/Estimate, Exp. 3: Full/ZS/Suggest, Exp. 4: Full/CoT/Suggest, Exp. 5: Limited/ZS/Suggest.

Summary

- Structure prompting (CoT) could significantly enhance specification quality
- Limiting detailed data access could lead to better theoretical reasoning by LLMs
- Claude excelled when assessed by AIC ➔ better in balancing goodness-of-fit and complexity
- ChaGPT better in terms of LL
- Only ChatGPT o3 is capable of both suggesting and correctly estimating MNLs

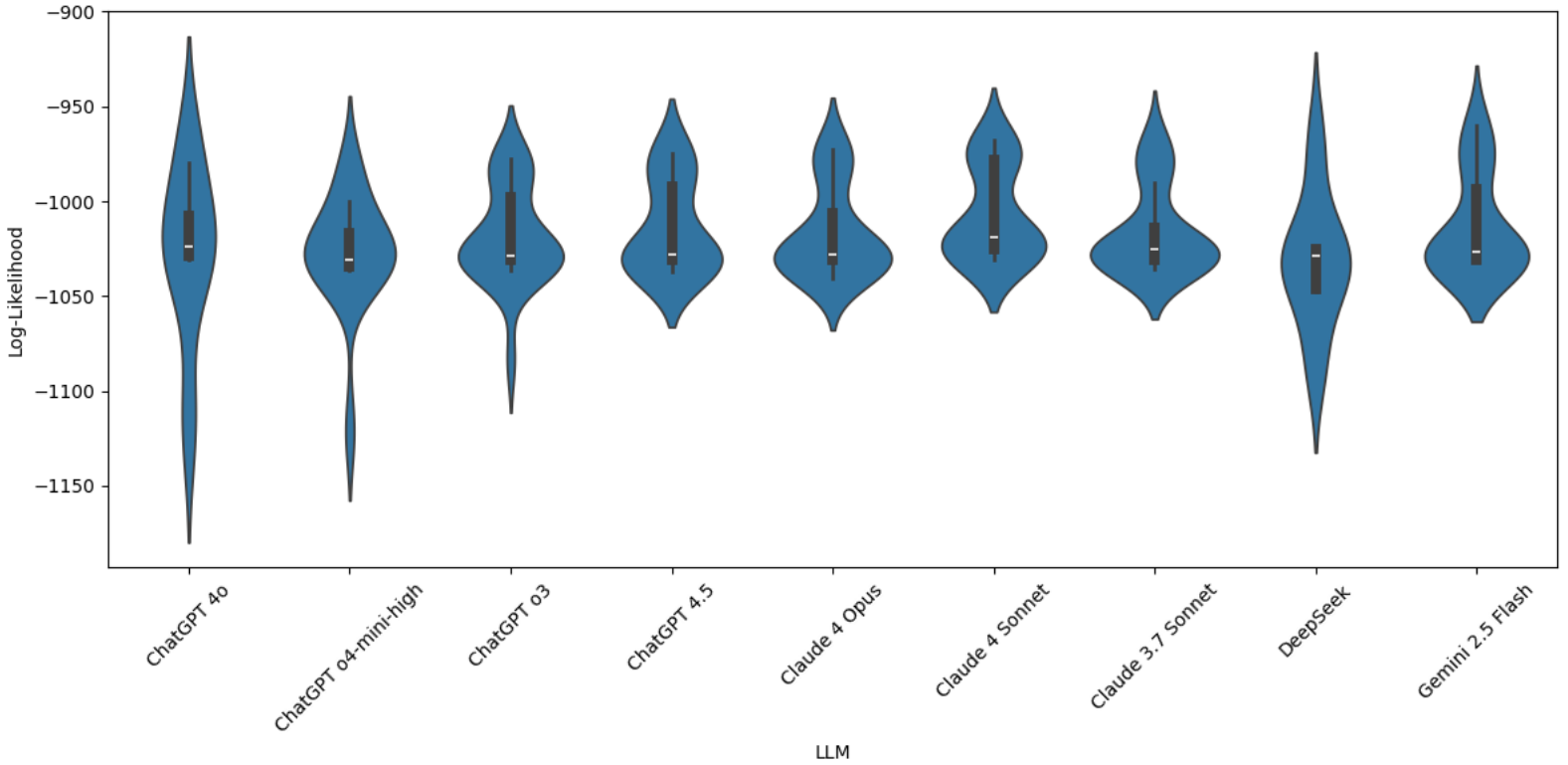
Overall Evaluation

Results – Specification Quality

LLM	Av. Nb of Spec.	Models Converged	Av. Nb of Vars	Av. Nb of Params	Generic Params	Alt-Spec. Params	ASC Included	Av. Nb of Socioeconomics	Av. Nb of Transformations	Av. Nb of Interactions
ChatGPT 4o	2.58	100%	4.00	10.75	23%	77%	58%	0.50	0.50	0.42
ChatGPT o4-mini-high	3.00	100%	3.53	7.80	47%	53%	93%	0.33	0.73	0.40
ChatGPT o3	2.77	100%	3.68	7.18	47%	53%	100%	0.27	0.59	0.59
ChatGPT 4.5	2.86	100%	3.93	9.00	35%	65%	100%	0.64	0.50	0.29
Claude 4 Opus	3.68	89%	3.68	7.74	43%	57%	100%	0.37	0.26	0.53
Claude 4 Sonnet	3.50	100%	4.67	9.33	36%	64%	100%	1.22	0.33	0.72
Claude 3.7 Sonnet	4.05	76%	3.90	9.81	33%	67%	100%	0.57	0.33	0.67
DeepSeek R1	3.35	76%	3.53	7.29	50%	50%	65%	0.18	0.29	0.47
Gemini 2.5 Flash	2.93	100%	3.50	10.71	32%	68%	100%	0.36	0.43	0.43

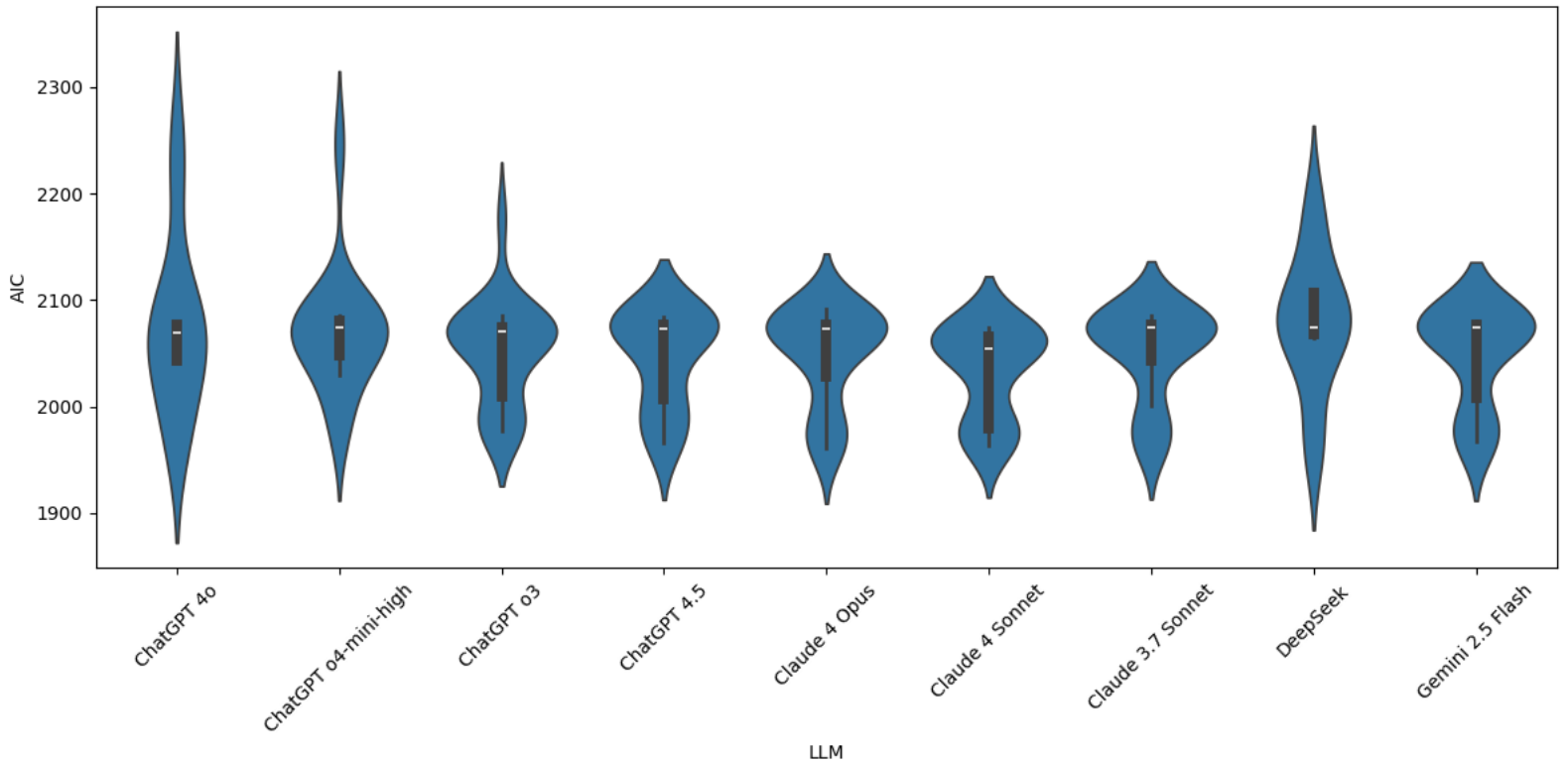
- All Claudes generated on average more specifications than all other LLMs
- No convergence issues for GPTs and Gemini
- Claude 4 Sonnet generated the most complex specifications: highest Nb of Vars (4.67), highest Nb of Socioeconomics (1.22), highest Nb of Interactions (0.72), and relatively high Nb of Params (9.33)
- ChatGPT o3 proposed simpler specifications
- DeepSeek suffered from convergence issues and omitting ASCs

Results – All Experiments



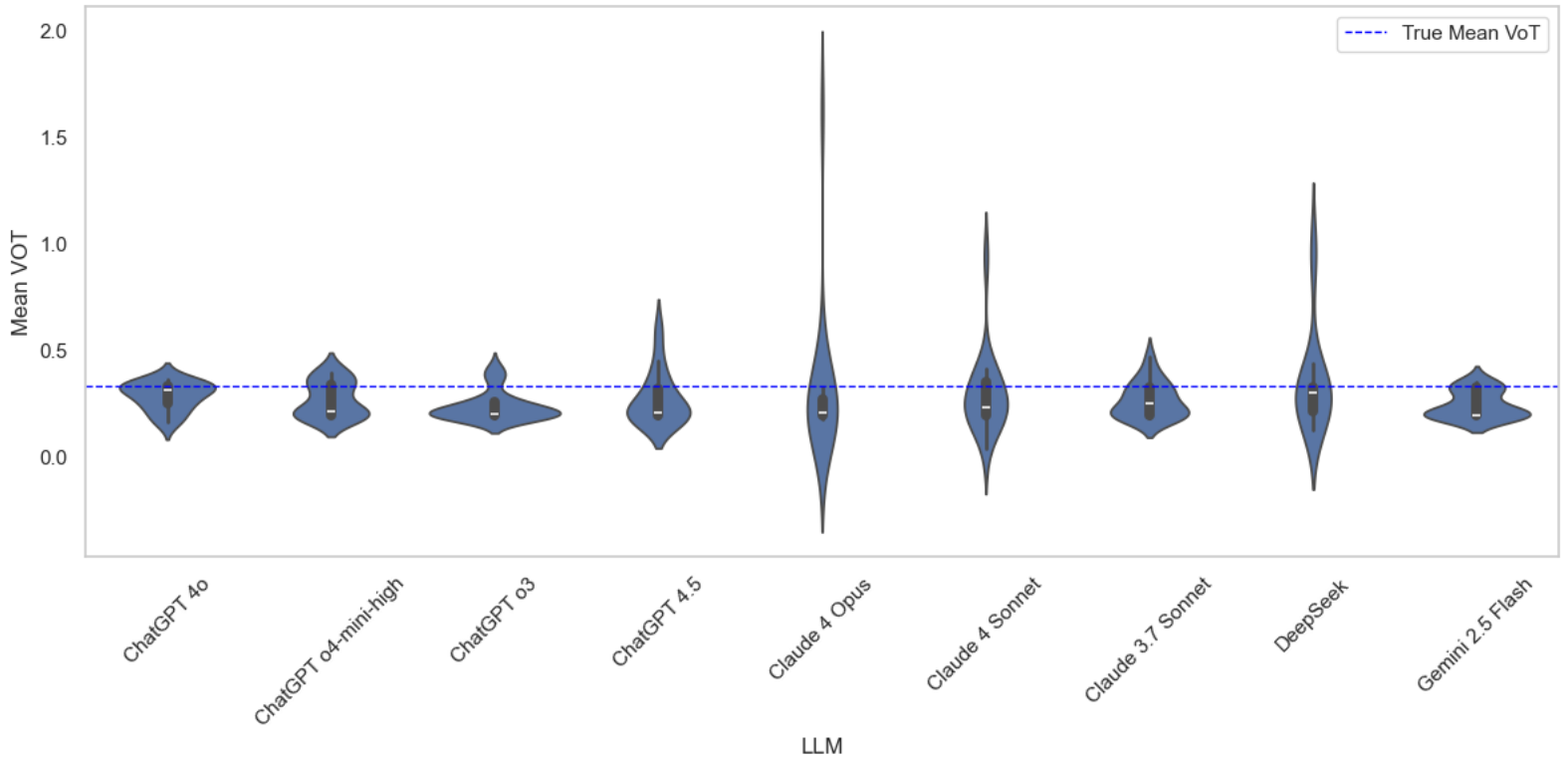
Models that did not converge are excluded

Results – All Experiments



Models that did not converge are excluded

Results – All Experiments



choice modelling centre

Conclusion

- LLMs show real potential in assisting discrete choice model development — especially in utility specification tasks
- Structured prompts (Chain-of-Thoughts) likely to improve model quality over Zero-Shot prompting
- Limited data access could sometimes lead to better theoretical reasoning and empirical performance
- GPT-o3 the only LLM capable of both specifying and correctly estimating MNL models end-to-end
- Claude models, especially Claude 4 Sonnet, produced the most complex and behaviourally rich specifications, albeit with occasional convergence issues

- Implication: AI cannot completely replace choice modellers… yet