Supplementary Information for "Identifying Collusive Corporate Communications at Scale with Language Models"*

Eduardo M. Azevedo† Joseph E. Harrington, Jr.‡ Ioan O. Rusu§

March 6, 2026

A Performance of Alternative Prompts and LLMs

We assess LLM responses on the benchmark sample against the human ratings for U-Haul and the airline industry. First, in Table 1 we compare different prompt versions, using the same gpt-4o-mini model, based on the first LLM response for each transcript-prompt pair (the exact text of the prompts will be available in the replication package). For robustness, we also repeat each transcript-prompt query another 10 times using the same model, and show the corresponding metrics based on the average of the 11 LLM responses (Table 2). We report the following metrics based on the corresponding confusion matrix for each prompt:

$$\text{Precision} = \frac{TP}{TP+FP}, \quad \text{Recall} = \frac{TP}{TP+FN}, \quad \text{Specificity} = \frac{TN}{TN+FP}, \quad \text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

We also compare LLM responses using different models from OpenAI that are suitable for our task. In Table 3 we compute for each model the same metrics we did for the prompt comparison, using the first LLM response for each transcript-model pair, and keeping the SimpleCapacityV8.1.1 prompt fixed. The models are grouped such that *modern* models are those that can be queried with a defined schema structure that the response will follow, the *reasoning-only* models are part of the most recent class of LLMs that produce responses following a more refined logic meant to improve quality, and the *older*

*Replication code available at `https://github.com/eduardomazevedo/collusion-llm`.

†Department of Business Economics & Public Policy, The Wharton School, University of Pennsylvania, 3733 Spruce Street, Philadelphia, PA 19104. eazevedo@wharton.upenn.edu, `https://eduardomazevedo.github.io/`.

‡Department of Business Economics & Public Policy, The Wharton School, University of Pennsylvania, 3733 Spruce Street, Philadelphia, PA 19104. harrij@wharton.upenn.edu, `https://joeharrington5201922.github.io/`.

§School of Economics, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104. ir79@drexel.edu.

models are a small selection of earlier LLMs which didn't allow imposing a response schema, so we used a version of the prompt that simply added to the original a concise description of the desired schema, following the guidance of OpenAI.

Note that the metrics we report here are just meant to provide a clearer way to compare quantitatively the responses across different prompts or models, but are not exhaustively assessing their performance. Our decision to use the `SimpleCapacityV8.1.1` prompt and gpt-4o-mini model for our large scale analysis of the full transcript sample was also largely guided by the other information the LLM responses contained (i.e. excerpts and reasoning). Moreover, specifically for the LLM selection, we considered the cost and token context size along with the quality of the responses, to ensure the feasibility of our large-scale analysis. We acknowledge that other prompt versions might be used, along with newer, more advanced models which are released frequently, to improve the LLM assessment for this analysis.

Table 1: Prompt Variants Benchmarking (First LLM Run)

Prompt	TP	FP	TN	FN	Precision	Recall	Specificity	F1
Basic Variants								
SimpleScoreV1	3	0	254	132	1.000	0.022	1.000	0.043
SimpleFlagV1	9	3	251	126	0.750	0.067	0.988	0.122
SimpleFlagScoreV1 (Score)	19	5	249	116	0.792	0.141	0.980	0.239
SimpleFlagScoreV1 (Flag)	22	11	243	113	0.667	0.163	0.957	0.262
Refined Variants								
PriceCapacityV1	27	7	247	108	0.794	0.200	0.972	0.320
PriceCapacityV2	26	8	246	109	0.765	0.193	0.969	0.308
PriceCapacityV3	60	16	238	75	0.789	0.444	0.937	0.569
PriceCapacityV4	84	53	201	51	0.613	0.622	0.791	0.618
PriceCapacityV5	41	12	242	94	0.774	0.304	0.953	0.436
PriceCapacityV6	44	9	245	91	0.830	0.326	0.965	0.468
SimpleCapacityV8	23	5	249	112	0.821	0.170	0.980	0.282
SimpleCapacityV8.1	34	7	247	101	0.829	0.252	0.972	0.386
SimpleCapacityV8.1.1	42	8	246	93	0.840	0.311	0.969	0.454
SimpleCapacityV8.2	33	10	244	102	0.767	0.244	0.961	0.371
SimpleCapacityV8.3	33	6	248	102	0.846	0.244	0.976	0.379
SimpleCapacityV8.4	32	8	246	103	0.800	0.237	0.969	0.366

Notes: Reported metrics use the Aryal et al. (2022) manual binary indicator for capacity discipline, our manual score rating for U-Haul transcripts, binarized at 50, and the LLM response score, binarized at 75. Where the LLM output includes both a score and a binary flag, we report the metrics on separate rows. All prompts are run using the gpt-4o-mini model on all benchmark sample transcripts.

Table 2: Prompt Variants Benchmarking (11 LLM Runs)

Prompt	TP	FP	TN	FN	Precision	Recall	Specificity	F1
Basic Variants								
SimpleScoreV1	0	0	254	135	0.000	0.000	1.000	0.000
SimpleFlagV1	2	0	245	111	1.000	0.018	1.000	0.035
SimpleFlagScoreV1 (Score)	2	1	253	133	0.667	0.015	0.996	0.029
SimpleFlagScoreV1 (Flag)	4	2	237	87	0.667	0.044	0.992	0.082
Refined Variants								
PriceCapacityV1	13	1	253	122	0.929	0.096	0.996	0.174
PriceCapacityV2	17	1	253	118	0.944	0.126	0.996	0.222
PriceCapacityV3	43	11	243	92	0.796	0.319	0.957	0.455
PriceCapacityV4	77	25	229	58	0.755	0.570	0.902	0.650
PriceCapacityV5	19	3	251	116	0.864	0.141	0.988	0.242
PriceCapacityV6	20	2	252	115	0.909	0.148	0.992	0.255
SimpleCapacityV8	7	1	253	128	0.875	0.052	0.996	0.098
SimpleCapacityV8.1	12	0	254	123	1.000	0.089	1.000	0.163
SimpleCapacityV8.1.1	17	1	253	118	0.944	0.126	0.996	0.222
SimpleCapacityV8.2	11	0	254	124	1.000	0.081	1.000	0.151
SimpleCapacityV8.3	12	0	254	123	1.000	0.089	1.000	0.163
SimpleCapacityV8.4	10	0	254	125	1.000	0.074	1.000	0.138

Notes: Reported metrics use the Aryal et al. (2022) manual binary indicator for capacity discipline, our manual score rating for U-Haul transcripts, binarized at 50, and the average LLM response score across the initial run and 10 repeated queries, binarized at 75. Where the LLM output includes both a score and a binary flag, we report the metrics on separate rows, and assign the flag value using a threshold of 0.5 for the averaged flag from the LLM responses. All prompts are run using the gpt-4o-mini model on all benchmark sample transcripts.

Table 3: LLMs Benchmarking (First LLM Run)

Model	TP	FP	TN	FN	Precision	Recall	Specificity	F1
Modern Structured Models								
gpt-4.1	11	0	254	124	1.000	0.081	1.000	0.151
gpt-4.1-mini	1	0	254	134	1.000	0.007	1.000	0.015
gpt-4.1-nano	4	1	253	131	0.800	0.030	0.996	0.057
gpt-4o	16	2	252	119	0.889	0.119	0.992	0.209
gpt-4o-mini	42	8	246	93	0.840	0.311	0.969	0.454
gpt-5	7	2	252	128	0.778	0.052	0.992	0.097
gpt-5-mini	7	2	252	128	0.778	0.052	0.992	0.097
gpt-5-nano	5	2	252	130	0.714	0.037	0.992	0.070
o1	3	0	254	132	1.000	0.022	1.000	0.043
o3	1	0	254	134	1.000	0.007	1.000	0.015
o3-mini	1	0	254	134	1.000	0.007	1.000	0.015
o4-mini	3	0	254	132	1.000	0.022	1.000	0.043
Reasoning-Only GPT-5.x Models								
gpt-5.1	3	0	254	132	1.000	0.022	1.000	0.043
gpt-5.2	1	0	254	134	1.000	0.007	1.000	0.015
gpt-5.3-chat-latest	0	0	254	135	0.000	0.000	1.000	0.000
Older Chat Models (JSON Prompt)								
gpt-4-turbo	4	1	253	131	0.800	0.030	0.996	0.057
gpt-3.5-turbo(*)	113	183	71	22	0.382	0.837	0.280	0.524

Notes: Reported metrics use the Aryal et al. (2022) manual binary indicator for capacity discipline, our manual score rating for U-Haul transcripts, binarized at 50, and the LLM response score, binarized at 75. Prompt is fixed at `SimpleCapacityV8.1.1`. For the model marked with (*), due to its context window size, 28/389 transcripts had to be trimmed (last approx. 7% of tokens) to fit into single LLM queries.

B Human Audit Sample Assessment

The transcript-level human-audit assessment file used in this project is available at this link. It includes the audited transcript identifiers, manual assessments, and supporting coding notes.

C Companies with Collusive Content from the Human Audit (Number = 71)

Table 4: Companies with Collusive Content from the Human Audit

Company	Primary Market Supplied
Air France-KLM	Passenger air transportation
AirTran	Passenger air transportation
Alliance Resource Partners	Thermal coal (energy coal)
American Airlines	Passenger air transportation
Ampco-Pittsburgh	Forged and cast metal components
Atwood Oceanics	Offshore oil drilling services
Avis Budget	Car rental services
Bharti Airtel	Mobile and fixed telecommunications
Boral	Construction materials (cement, aggregates)
Cameco	Uranium mining and supply
Casella Waste	Solid waste collection and disposal
Cementir	Cement production
CEMEX	Cement and ready-mix concrete
Charles River Laboratories	Preclinical research services for pharmaceuticals
Cleveland-Cliffs	Iron ore and steel production
CNX Gas	Natural gas production
Coca-Cola Bottlers Japan	Non-alcoholic beverages (soft drinks)
Companhia Siderúrgica Nacional	Steel production
Continental Resources	Oil and gas exploration and production
Corning	Specialty glass and materials
Daqo New Energy	Polysilicon for solar panels
Delta Airlines	Passenger air transportation
Embraer	Commercial and regional aircraft manufacturing
Essity	Hygiene and tissue paper products
Expand Energy	Oil and gas exploration and production
Ferroglobe	Silicon metal and manganese alloys
Frontier	Passenger air transportation
Glatfelter	Specialty paper products

Company	Primary Market Supplied
Glencore	Commodity trading and mining
Goldcorp	Gold mining
Gol Linhas	Passenger air transportation
Green Plains	Ethanol and biofuels
Gulfport Energy	Natural gas production
Harley-Davidson	Motorcycles
Huntsman	Specialty chemicals
ICL	Fertilizers and specialty minerals
Italcementi	Cement production
JinkoSolar	Solar panels
Kirby	Inland marine transportation (tank barges)
Micron Technology	Memory semiconductors
Norske Skogindustrier	Newsprint and publication paper
O-I Glass	Glass containers
Oladapco	Oil products distribution (Nigeria)
Ovintiv	Oil and gas exploration and production
PGS	Seismic data services for oil and gas
Pilgrim's Pride	Poultry products
ProFrac Holding	Hydraulic fracturing services
Prosafe	Offshore accommodation rigs
PT XLSMART Telecom	Mobile telecommunications
Reliance Communications	Telecommunications
Sappi	Paper and pulp
Seadrill	Offshore oil drilling
Seagate Technology	Data storage (hard disk drives)
Six Flags Entertainment	Amusement and theme parks
Smithfield Foods	Pork and processed meat products
STEP Energy	Oilfield services
Superior Energy	Oilfield services
Suzano	Pulp production
Telecom Italia	Telecommunications
Teva Pharmaceuticals	Generic pharmaceuticals
Tidewater	Offshore marine services
Trip.com Group	Online travel booking platforms
Ultra Petroleum	Natural gas production
United Airlines	Passenger air transportation
US Airways	Passenger air transportation
Vodafone	Mobile telecommunications
Waste Management	Solid waste collection and disposal
Western Digital	Data storage devices

Company	Primary Market Supplied
WestRock Paper and Packaging	Paperboard and packaging
WH Group	Pork and meat products
Wienerberger	Clay bricks and building materials

Notes: For the 301 transcripts with the highest LLM-validated scores used in the human audit, these are the companies whose transcripts were found to have collusive content. The "primary market supplied" was generated by ChatGPT 5.2 and then verified by research assistant Vivian Zhang.

D Categorization of Collusive Content

Industry reduced supply or capacity and references competitors' conduct or the industry's interests in a manner supportive of such conduct (40 transcripts)
508031, 597966, 505840, 508027, 419393, 352683, 508036, 352685, 392548, 481676, 481654, 602956, 419394, 481658, 481659, 174683, 416422, 524742, 392547, 481690, 427773, 591146, 597969, 259613, 363812, 381391, 185218, 389910, 634653, 260640, 346016, 611671, 569063, 261028, 348864, 340019, 508199, 215550, 190662, 445956

Industry needs to reduce supply or capacity or show discipline (30 transcripts)
183145, 418022, 267433, 244726, 363814, 519067, 174682, 427772, 627577, 406151, 481652, 325885, 260639, 611639, 494779, 508184, 381486, 617313, 352688, 342599, 199854, 227682, 399246, 260628, 199856, 436918, 260643, 228511, 327069, 617015

The industry has shown or will show supply or capacity or price discipline (12 transcripts)
196851, 392562, 227685, 399267, 396666, 445959, 508185, 264077, 309604, 558080, 396668, 431603

Industry needs to raise prices (12 transcripts)
250287, 408614, 215003, 348098, 558083, 558091, 321896, 300093, 375388, 363809, 305524, 324730

The firm is being or will be less aggressive in pricing and references competitors' conduct in a manner encouraging such conduct (8 transcripts)
558086, 377924, 375653, 309638, 223177, 526439, 631460, 321901

Miscellaneous (7 transcripts)
States an industry objective that will tend to lead to higher prices [325064, 408622]
Industry will limit capacity or supply or not overproduce [252937, 445983]
All firms raised prices and pricing strength is expected to continue [257344]
Firm welcomes consolidation, whether it's done by us or others as it will add to discipline and make a better market for all of us [360150]
Firm will not oversupply and hopes its dealers stop competing with themselves [548932]

E Content Analysis of False Positives

Out of the 301 LLM-validated transcripts with the highest scores, 192 were classified by the human auditor as false positives. Of those, 29 of them lacked any relevant content and have been interpreted as hallucinations. The other 163 contain some content which, in principle, could be part of a collusive message. Specifically, the transcripts refer to the firm's competitive conduct or the state of market competition though did not connect it to the conduct of competitors or the industry and thereby fell short of facilitating coordinated conduct. Here we describe the most common content that appeared in the LLM's excerpts and appeared to be the reason for the LLM assigning it a high score.

67 transcripts referred to a firm reducing, restricting, or managing its supply or capacity in a manner consistent with higher prices or avoiding lower prices. However, it did not say or suggest that this conduct is something the industry should or will or needs to do.

> "Our intention [is] to maintain a disciplined sales approach. Rather than selling at material decrease prices. We implemented a strategy to hold back product [and] resulting in a significant buildup of the inventories."
> [OCI N.V. | Earnings Calls | 24 May 2019]

> "We will continue to manage our customer needs and will not restart capacity on a whim just to add tonnage to the spot market. That would not be good for anyone. Value over volume is a simple philosophy that we will carry on into the future, and it's why you are not going to hear me talk about capacity utilization or even market share."
> [Cleveland-Cliffs Inc. | Earnings Calls | 25 Feb 2021]

> "We stand by our strategy of not forsaking value for volume and protecting our profitability with continued actions on our cost structure [and we] cut our production capacity."
> [PIRELLI & C. Società per Azioni | Earnings Calls | 12 Nov 2012]

> "We go down by elevator and we go up by stairs. So to avoid this huge volatility, we are aiming . . . to reduce capacity."
> [Fibria Celulose S.A. | Earnings Calls | 26 Apr 2017]

By noting its leadership position or forecasting constrained industry supply, some announcements do get close to suggesting that firms coordinate on reducing supply or capacity.

> "Our first step is to stop the significant price declines. The most important tool that we have to accomplish this is to reduce capacity to get glass supply in line with market demand. When you're more than half the market, that really adds up."
> [Corning Incorporated | Analyst/Investor Day | 03 Feb 2012]

"And until now, we have not heard anyone else announcing any capacity enhancements in the short term. None of them is increasing capacity."
[HEG Limited | Earnings Calls | 16 Aug 2021]

"Additional capacity reductions can be expected [and] we forecast an extended period when the … industry operates a smaller, more profitable footprint. Our outlook is for lumber pricing to gradually increase."
[Conifex Timber Inc. | Earnings Calls | 14 Aug 2019]

48 of the transcripts refer to higher prices - whether noting having raised prices or expressing a need to raise prices - or to pricing power. Such a statement might be made in the context of reducing supply.

"We reduced our … production capacity by more than 20% [and] we have already announced another 20% price hike. We have started to see the sign of improving pricing trend."
[AGC Inc. | Earnings Calls | 06 Aug 2012]

"We're going to... be price stewards. We don't chase volume with price."
[Rayonier Advanced Materials Inc. | Analyst/Investor Day | 10 Oct 2023]

"We have taken a call … to increase prices of cement to help manage the rising cost of production."
[UltraTech Cement Limited | Earnings Calls | 18 Oct 2021]

With a reference to competitors reducing capacity in the context of an emphasis on raising prices, this announcement is getting close to having collusive content.

"In the short term, there's only one priority, and that's price increases and price increase and price increases. We have good capacity utilization. We have strong brands. We always go for pricing every time. Many competitors, especially in the tissue area, announced that they were shutting down paper machines due to this is looked at an Italian player ... and I think we'll see more of that."
[Essity AB (publ) | Company Conference Presentations | 10 Mar 2022]

In several transcripts, a firm announced it was striving to avoid lowering prices.

"We concluded that lowering price to gain share in a market like we are experiencing this year was not the right strategy."
[Avis Budget Group, Inc. | Earnings Calls | 03 Nov 2009]

"We don't want to underprice and overshoot on demand. We keep some of the exclusive capacity in order to be able to sell at high prices. We want to be profitable [even] if we lost a couple of market share percentage points."
[TUI AG | Earnings Calls | 11 May 2022]

"[We are] idling significant chunks of our assets and slashing our participation in weak markets. We're not just going to participate in those weak markets because it's a chase down into the mud."
[Olin Corporation | Earnings Calls | 27 Oct 2023]

A firm indicated a strategy to strengthen pricing power.

"We deliberately allocate inventory to fiscal years and regions to ensure we maintain apparent scarcity and, therefore, protect our brand health and protect our pricing power."
[Treasury Wine Estates Limited | Earnings Calls | 16 Aug 2018]

"We're in the early stages of having a terrific opportunity over the next 5 years to be in a position where we can increase the returns at our disposal facilities, primarily through price. We're one of the only players with a little bit of excess capacity left."
[Casella Waste Systems, Inc. | Company Conference Presentations | 14 Jun 2018]

A firm described how it is being more disciplined and less aggressive which, if it had expressed competitors acting or needing to act in a similar manner, would have encompassed collusive content.

"We do not want to accelerate a rapid decline in price of the products. We will only ramp it up very moderately when we start production because we do not want to accelerate a collapse in price. We want to be more disciplined in our approach."
[LANXESS Aktiengesellschaft | Earnings Calls | 06 Nov 2014]

"We are not chasing market share. We are focusing on quality customers and then driving spend on our network by our existing customers."
[Dialog Axiata PLC | Earnings Calls | 18 Feb 2025]

"We have ... implemented Microchip's disciplined pricing process ... to ensure that we are competitive and don't chase bad business in the pursuit of profitless prosperity."
[Microchip Technology Incorporated | Earnings Calls | 08 Aug 2016]

In this case, the firm is expressing an aspiration for the industry which is getting close to collusive content.

"We would like to drive the industry in a direction where we can move back to more positive margins."
[Total Access Communication Public Company Limited | Earnings Calls | 10 Feb 2015]

A firm noted that it has market power or what the sources of market power are. While in one case it does mention reduced industry capacity being a source, it is of the form of a factual observation.

> "Supply is having a harder time to react to demand [and] that's [where] we're getting the pricing power."
> [Prologis, Inc. | Company Conference Presentations | 10 Sep 2019]

> "When our actual sales pace exceeds our budgeted sales pace, we try to push prices."
> [Hovnanian Enterprises, Inc. | Earnings Calls | 09 Sep 2013]

> "We expect to continue to implement favorable pricing actions ... due to capacity constraints in our industry. We are operating today in a very strong demand environment without enough industry capacity to meet that demand and expect that to continue into the future."
> [Carpenter Technology Corporation | Special Calls | 18 Feb 2025]

There are some transcripts referring to a reduced industry supply or capacity but only in the context of stating it as a fact. Still, the announcements are getting close to content that may encourage competitors to reduce supply or capacity.

> "This moment of rationalization of supply is new in the history of the Brazilian civil aviation industry. We wouldn't like to start or restart price war in the Brazilian market. Our strategy of reducing capacity is indeed in place and it's showing results."
> [Gol Linhas Aéreas Inteligentes S.A. | Earnings Calls | 14 May 2013]

> "[There is] limited new barge construction in the industry and many units going in for maintenance. So we should have a net decline this year in supply. Nobody is really panicking and trying to go out and build any new equipment right now."
> [Kirby Corporation | Earnings Calls | 01 Feb 2024]

In 14 transcripts, the firm refers to industry consolidation by noting that it has been consolidating, needs to consolidate, or will consolidate. In some instances, the firm attributes consolidation to too much supply or capacity or too many firms. In one case, they do seem to be facilitating such consolidation.

> "We expect the industry will consolidate to approximately 5 Pan-India players. Once that happens, the industry will further see tariff hardening, with pricing power returning."
> [Reliance Communications Limited | Earnings Calls | 10 Feb 2014]

> "In order to really improve value creation, we need consolidation and further restructuring. There's a need for further consolidation and restructuring in this industry to be able to really actually take the costs down and balance the markets as well."
> [UPM-Kymmene Oyj | Earnings Calls | 03 Aug 2011]

In summing up, the LLM identified instances in which a firm's message was consistent with less competition in that it referred to reducing supply or capacity or raising or at least not lowering prices. To facilitate collusive conduct, a firm needs to link such content to industry conduct in a manner that encourages similar conduct by competitors. The LLM failed in not verifying that such a link was present in the firm's message. Consequently, the LLM had a high false positive rate.