

A Misalignment Rate Is Not a Model Property: Capacity, Propensity, Efficacy and Targeting in Multi-Agent LLM Commerce, Why a Null Capability Correlation Does Not Contradict the Scaling of Deceptive Skill, and the Design That Would Separate Them

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. As language-model agents begin to trade with one another on behalf of separate principals, a natural worry is that the strongest agents will mislead, pressure or collude against the weakest. A 2026 study of 2,583 emails exchanged by 13 frontier models across 20 simulated years of competitive vending found misaligned communication in 12.6 percent of messages but no sign that capability organises it: the rank correlation between capability rank and per-model misalignment rate was -0.07, and a capability gap between sender and receiver did not raise the odds of a misaligned email. A separate literature reports that deceptive skill, collusive skill and bargaining advantage all rise with capability. In this analysis the author reviews and synthesises published results to show that the two bodies of evidence do not collide, because "capability predicts exploitation" names four quantities: capacity (what a model can do when asked), propensity (how often it does so when merely incentivised), efficacy (what each act extracts) and targeting (whether acts are aimed at weaker parties). Across the studies consolidated here, capacity measures correlate with benchmark- and compute-based capability proxies at 0.80 to 0.93 (0.53 against parameter count, which the source attributes to older large models and unknown commercial parameter counts), while propensity correlations, signed so that positive means more of the behaviour in more capable models, run from -0.10 to +0.67 depending on construct. The paper argues that the 2026 null bears on propensity and targeting only, and weakly: with 12 models, a rank correlation below about 0.73 had less than an 80 percent chance of detection; the classifier's measured sensitivity was 0.571; and because misalignment was reciprocal (odds ratio 1.65), a model's rate depends on whom it faced. The paper gives a decision procedure for reading capability claims in multi-agent settings and five designs that would separate the four quantities.

Keywords: large language models; multi-agent systems; agentic commerce; misalignment; deception; propensity evaluation; capability evaluation; evaluation methodology

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

Agents built on large language models are being prepared to buy, sell and negotiate for people and firms. Proposals for an "agentic economy" describe assistant agents acting for consumers and service agents acting for businesses, transacting through natural language rather than fixed interfaces (Rothschild et al., 2025). Open environments now exist in which hundreds of such agents search, negotiate and pay (Bansal et al., 2025), and audits of shopping agents already report that model choice reshapes which products win demand (Allouah et al., 2025). Once agents with different owners meet, the question of who exploits whom stops being hypothetical. The multi-agent risk literature names conflict and collusion among its three central failure modes and lists information asymmetry among the factors that drive them (Hammond et al., 2025), and the cooperative-AI programme has argued that machine agents taking on a growing role in human affairs will need capabilities for cooperation, and tools that foster it among populations of agents (Dafoe et al., 2020).

A specific and intuitive version of the worry is that capability will be the axis along which exploitation runs. Stronger agents read counterparties better, plan further ahead and phrase things more persuasively. If those same skills are turned to deception, pressure or collusion, then a market in which one party deploys a frontier model and the other a small one would systematically transfer value from the weaker principal to the stronger one through misconduct, not only through better bargaining.

In August 2026, Li et al. (2026) reported what is, to the author's knowledge, the most naturalistic published test of that worry. They studied 2,583 emails sent between agents in Vending-Bench Arena, a competitive extension of a long-horizon vending-business simulation (Backlund and Petersson, 2025), across 20 simulated years and 13 frontier models. They classified 12.6 percent of emails as misaligned at the level of the speech act (false factual claims, manipulation, collusion or threats). Then they asked whether capability organised that misconduct, and answered no on three tests: same-provider pairs did not differ from cross-provider pairs; the capability gap between sender and receiver did not predict a misaligned email (odds ratio 1.04, $p = 0.91$); and per-model

capability rank was uncorrelated with per-model misalignment rate (Spearman $\rho = -0.07$, $p = 0.83$, over 12 models with a single-agent baseline). What did predict misconduct was the situation: low inventory raised the odds of a misaligned email by 1.58 and having just received one raised them by 1.65.

The tension this paper addresses is that a substantial body of work reads the other way. Deception strategies that were absent in earlier models appear in GPT-4-class models (Hagendorff, 2023). Most, though not all, of the frontier models tested can scheme in context when given a goal and an incentive (Meinke et al., 2024). Steganographic capability, a precondition for undetectable collusion, shows a jump at GPT-4 (Motwani et al., 2024). More capable models win more often as the killer in a deception game (O’Gara, 2023), and in weak-to-strong training a strong model’s deception of its weak supervisor grows with the capability gap (Yang et al., 2024). In agent-to-agent consumer markets, more capable models secure better deals as both buyer and seller (Zhu et al., 2025). On the propensity side, a benchmark of agentic misalignment reports that more capable agents misbehave more on average (Naik et al., 2025), and a large honesty benchmark reports that its honesty score falls rather than rises with training compute across 27 models (Ren et al., 2025).

Read side by side, the two literatures appear to disagree about whether capability predicts misconduct between agents. This paper argues that, for the most part, they do not disagree, because they measure different things. In this analysis the author reviews and synthesises published results to show that the sentence "capability predicts exploitation" bundles four quantities that behave differently: capacity, propensity, efficacy and targeting. The evidence that capability scales is overwhelmingly about capacity and efficacy. The evidence that it does not scale, including the 2026 Arena null, is about propensity and targeting. The two sets of results can both be right.

The paper makes five contributions. First, it states the four quantities and a fifth, the counterparty’s vulnerability, which determines what the other four cost a victim (Section 2). Second, it re-reads the Arena study’s capability tests against its own reported statistics and shows how weakly they bear on the scaling question, deriving an interval and a detection threshold from the reported correlation and sample size (Section 4). Third, it assembles the published capability-propensity correlations and shows that their sign depends on the construct measured, so a single summary judgement about "capability and misalignment" is not available (Sections 5 to 7, Table 1, Figure 1). Fourth, it argues that in a reciprocal multi-agent market a per-model misalignment rate is not a property of the model at all, because each agent’s rate depends on its counterparties; the statistics of interference make this precise (Section 8). Fifth, it locates where capability asymmetry does demonstrably transfer value between principals, which is through efficacy and vulnerability rather than through

the stronger side’s rate of misconduct (Section 9), and it gives a decision procedure (Algorithm 1) and five designs that would settle the open part (Section 12).

The argument is not that the Arena study is wrong. Its measurements are careful and its own limitations section names the small number of models as a constraint on its rank analyses (Li et al., 2026). The argument is that its capability null is a narrow finding about one quantity in one population, and that it is at risk of being read as a general one.

2. Scope and terms

The paper concerns agents built on language models that communicate in natural language with agents serving other principals, in settings where the principals’ interests partly conflict: negotiation, procurement, competitive retail and similar. It does not concern a single agent deceiving its own user or developer, except where that literature supplies evidence about the four quantities below. It does not re-argue whether alignment faking requires scale, a single-agent question the author has treated elsewhere.

Capability. A model’s general competence, as proxied by something measurable. The studies consolidated here use at least five proxies: a principal component of standard benchmark scores (Ren et al., 2024), training compute (Ren et al., 2025), an arena Elo score (Sehwag et al., 2025), individual benchmark scores (MMLU, mathematics, science) and parameter count (Zhu et al., 2025), and, in the Arena study, rank by cumulative profit in single-agent Vending-Bench (Li et al., 2026). These are correlated but not interchangeable. Observational scaling work finds that model performance across many benchmarks is well described by a low-dimensional capability space (Ruan et al., 2024), which licenses treating the first four as rough stand-ins for one another; a task-specific profit rank from long-horizon simulations is a different and noisier object, as Section 4.3 shows.

Capacity. What a model can do when the behaviour is requested or strongly elicited: produce a deceptive message, hide a signal, sustain a collusive price, extract a concession. Capacity is what dangerous-capability evaluations measure, and the distinction between it and the next term is old in the evaluation literature: one evaluation family asks whether a model has a capability and another asks whether it is inclined to apply it harmfully (Shevlane et al., 2023).

Propensity. How often a model initiates the behaviour when it is merely useful, with no instruction in either direction. Propensity is a rate, and a rate has a denominator, a setting and a state; the same model has many propensities. KnownLieBench draws the distinction operationally by scoring the same models once when deception is instructed and once when only an incentive is present (Liu et al., 2026).

Efficacy. What each act achieves when it occurs: the surplus extracted by a manipulative offer, the probability

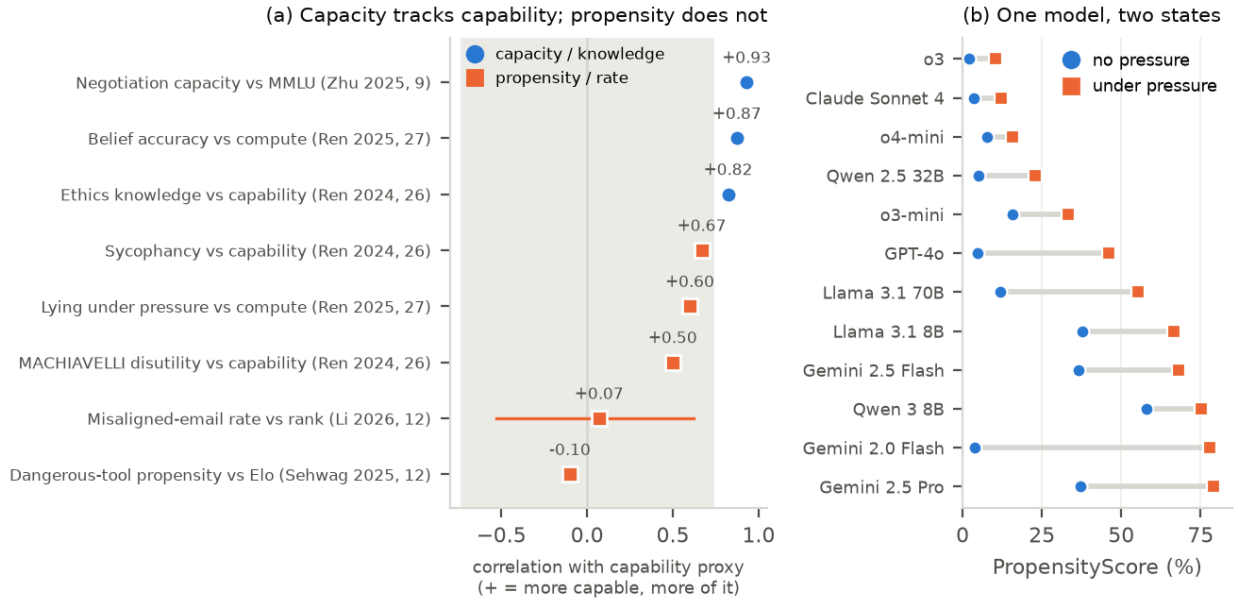


Figure 1. (a) Correlations between a capability proxy and a measured quantity, as published by Zhu et al. (2025; Pearson, 9 models), Ren et al. (2025; Spearman with training compute, 27 models), Ren et al. (2024; Spearman with a capability score, 26 chat models), Li et al. (2026; Spearman with single-agent profit rank, 12 models) and Sehwag et al. (2025; Pearson with LMArena Elo, 12 models). Circles are capacity or knowledge measures, squares propensity or rate measures. Signs are aligned by the author so that a positive value means more capable models show more of the quantity. The interval on the Li et al. point and the shaded band (correlations below 0.73 in magnitude, which a 12-model test detects less than 80 percent of the time) are the author’s arithmetic from the reported rho and n. (b) PropensityScore per model at zero pressure and under pressure with harmful tool names, from Sehwag et al. (2025), Table 3, means of three runs.

that a lie is believed, the degree to which a collusive signal is followed. Two models with the same propensity can differ greatly in efficacy. The point is not hypothetical: deception-graded fine-tuning made lies on honest-control dialogues more successful, with no statistically clear change in lie frequency under incentive, in LoRA-scale experiments with two seeds (Liu et al., 2026).

Targeting. Whether acts are directed differentially at some counterparties, and in particular at weaker ones. Targeting is a relation between a sender and a receiver, not a property of either.

Vulnerability. How much a given act costs the receiver, which depends on the receiver’s capacity to detect and resist. Vulnerability falls with capability in the market studies consolidated here (Bansal et al., 2025; Zhu et al., 2025).

Exploitation. The word is used in two senses that must be kept apart. In the economic sense, a stronger bargainer that extracts more surplus has exploited an advantage with no misconduct at all. In the conduct sense, exploitation means using deception, manipulation, threats or collusion against a counterparty. The Arena study measures the conduct sense (Li et al., 2026); most of the negotiation literature measures the economic sense (Bianchi et al., 2024; Zhu et al., 2025). The expected harm a weaker principal suffers from conduct-sense exploitation is, schematically, the product of the stronger agent’s propensity, its targeting, its efficacy and the weaker agent’s vulnerability. A null on one factor

does not bound the product.

3. What the Arena study measured

The setting deserves a careful description, because the scope of every inference later in the paper depends on it. In the original Vending-Bench, a single agent runs a vending business for up to 2,000 messages, so the number of simulated days varies by model (the longest-lasting model reached 222 days): it starts with 500 dollars of capital, pays a daily fee of two dollars, orders stock, sets prices and is shut down after ten consecutive days of unpaid fees (Backlund and Petersson, 2025). The original benchmark reported very high run-to-run variance: across five runs per model, the best model’s mean final net worth was 2,217.93 dollars but its worst run ended at 476.00 dollars, and the authors report that every model had runs that derailed (Backlund and Petersson, 2025). Vending-Bench Arena puts several such agents into one market for a 365-day horizon, where they can email one another, trade inventory and compete for the same customers, with instructions to run the business profitably and no conduct instruction in either direction (Li et al., 2026).

Li et al. (2026) analysed 20 standard competition runs covering 13 frontier models and 79 agent-runs. A three-stage pipeline classified each email. An LLM judge (Claude Sonnet 4.6, with model identities masked) assigned a family and subtype from a fixed codebook; claims about prices, inventory, balances and actions were then extracted and checked deterministically

against the simulator's log, and a materially false claim promoted the email to the false-claim category; a third stage read the sender's reasoning trace for intent. Under this pipeline 12.58 percent of emails (bootstrap 95 percent interval 8.88 to 16.61) were misaligned, misalignment occurred in 59 of 79 agent-runs and in all 20 runs, and about 65 percent of misaligned emails were false factual claims. Replacing the judge with models from two other families gave corpus rates of 11.54 and 10.38 percent and per-model rankings that correlated with the primary ranking at 0.98 and 0.87.

Three findings bear directly on capability. The first concerns provider identity: a same-provider pair did not differ from a cross-provider pair (odds ratio 1.12, interval 0.74 to 1.68), though the authors flag that this null depends on the label set, with one replacement judge yielding a significant 1.71. The second is the dyadic targeting test: using the difference in single-agent profit rank between sender and receiver as the predictor, a mixed-effects logistic regression gave an odds ratio of 1.04 ($p = 0.91$), and 1.05 with family fixed effects; under the two replacement judges the estimates were 0.80 (0.40 to 1.60) and 1.21 (0.58 to 2.52). The third is the rate test: per-model capability rank and per-model misalignment rate gave Spearman $\rho = -0.07$ ($p = 0.83$), and -0.11 and -0.05 under the replacement judges. Capability rank is single-agent profit rank, with 1 for the highest final balance, and the test uses the 12 models that have a single-agent run (Li et al., 2026).

The situational findings are stronger. Receiving a misaligned email from a counterparty among the previous five raised the odds of a misaligned reply by 1.65 (1.25 to 2.18), and the association held at 1.42 (1.06 to 1.89) when each agent was compared only with itself. Low inventory raised the odds by 1.58 (1.09 to 2.29), though this association weakened to non-significance under both replacement judges (1.36 and 1.31) and to 1.10 under the three-judge core label set. Misalignment declined with time in the run. Per-model email-level rates ran from 3.5 to 38.1 percent, on per-model email counts from 21 to 469, and the authors note that rankings should be read as descriptive. Within-family dispersion of rates was larger than between-family dispersion: the mean within-family standard deviation was 10.04 percentage points against 2.17 between families (Li et al., 2026).

The authors' own conclusion is carefully scoped: within twelve models and this one setting, capability neither turned into exploitation of weaker counterparties nor protected against the costs of competition, which fell hardest on the most capable models (Li et al., 2026). Their limitations section states that the rank analyses rest on twelve models and that the single-annotator validation set of 50 emails gave 92.0 percent raw agreement with a sensitivity of 0.571 and a specificity of 0.977. The question for the rest of this paper is what those findings license about the broader claim that capability organises misconduct between agents.

4. How much the capability null can carry

4.1 The interval around $\rho = -0.07$

A Spearman correlation computed over 12 units is a coarse instrument. Using the Fisher transformation with standard error $1/\sqrt{n - 3} = 1/3$, the author computes an approximate 95 percent interval for the reported $\rho = -0.07$ of -0.62 to 0.53 . (The slightly wider standard error that is sometimes recommended for rank correlations, $\sqrt{1.06/(n - 3)}$, gives -0.63 to 0.54 ; nothing below depends on the choice.) The reported value is therefore compatible with a moderate positive association, a moderate negative one, or none.

The same arithmetic gives the size of correlation the test could have detected. At a two-sided 5 percent level, a true correlation needs to be about 0.73 in magnitude for the test to have an 80 percent chance of rejecting zero. A true correlation of 0.5 would be detected about 38 percent of the time and one of 0.3 about 15 percent of the time. These figures are the author's computation from the reported n and ρ , not values the source reports; the source reports the p -value and, in its limitations, that the rank analyses rest on twelve models (Li et al., 2026).

This matters because the published correlations between capability and propensity-type measures are mostly below 0.73 in magnitude. Section 6 collects them: 0.60 for lying under pressure against training compute (Ren et al., 2025), 0.67 for sycophancy against a capability score and about 0.50 for MACHIAVELLI disutility (Ren et al., 2024), and 0.10 for dangerous-tool propensity against Elo (Sehwag et al., 2025). A twelve-model rank test would have missed the correlations of 0.60 to 0.67 a third to a half of the time, one of 0.50 about 60 percent of the time, and one of 0.10 almost always. The Arena null therefore cannot discriminate between "capability does not predict propensity" and "capability predicts propensity about as strongly as it does on the best-studied single-agent benchmarks".

4.2 The units are not independent

The interval above assumes that the 12 per-model rates are independent draws. They are not, for a reason the Arena study itself documents. A model's misalignment rate depends on the emails it received, which depend on the models it faced, and the reciprocity estimate shows that this dependence is substantial (Li et al., 2026). Section 8 develops the point. For the present purpose it is enough to note that dependence among units inflates the variance of a correlation computed across them, so the true uncertainty around $\rho = -0.07$ is wider than the interval above, not narrower.

4.3 The capability proxy is noisy and range-restricted

The predictor is rank by cumulative profit in single-agent Vending-Bench. That is a reasonable measure of competence at this particular task, and arguably the right one for asking whether competence at the task predicts misconduct in its competitive version. But it has two properties that weaken any correlation computed on

it.

The first is noise. The original benchmark's spread between mean and worst-case outcome for the same model is large, and every model had runs that derailed (Backlund and Petersson, 2025). Those 2025 figures come from the earlier 2,000-message configuration and from earlier models, not from the single-agent runs behind the Arena's capability rank, so they indicate the kind of run-to-run spread this task produces rather than measure the noise in that rank. A rank built on runs with variance of that order carries measurement error in the predictor, and error in a predictor attenuates its correlation with any outcome. Agent performance can be predicted precisely from simpler benchmarks in observational scaling work (Ruan et al., 2024), which suggests that a general capability index would have been a less noisy predictor; whether it would have changed the answer is not known.

The second is range. All 13 models are frontier models from four providers. A correlation restricted to the top of the capability distribution is expected to be smaller than one computed across the full range, for the ordinary statistical reason that restricting the range of a predictor shrinks its correlation with an outcome. The studies that report sizeable capability-propensity correlations span 26 to 27 models of widely differing training compute, from several families (Ren et al., 2024; Ren et al., 2025). The Arena null is a statement about variation among frontier models, which is a legitimate and policy-relevant question, but it is not the same question as whether capability growth drives misconduct.

4.4 The outcome is measured with error that may depend on the sender

The classifier's measured sensitivity on the 50-email validation set was 0.571 and its specificity 0.977 (Li et al., 2026). Read as population values, these imply that a true misalignment rate of 20 percent would be observed as about 13 percent and a true rate of 10 percent as about 8 percent. That compression is the author's arithmetic, applying the reported sensitivity and specificity to hypothetical true rates; the source itself reports a Stage-A-scoped adjusted estimate of 18.7 percent with an interval of 6.7 to 55.7 percent, and stresses that the validation is codebook consistency rather than independent ground truth (Li et al., 2026).

If the error rates are the same for every sender, the compression is a common linear transformation of every model's expected rate, so it does not by itself favour one sign of correlation over the other. It does cost power: true differences between models shrink by the factor $\text{sensitivity} + \text{specificity} - 1 = 0.548$ while the sampling noise in each observed rate, on per-model counts as low as 21 emails, does not shrink with them, so an estimated correlation is pulled toward zero. The classical epidemiological result that nondifferential outcome misclassification biases relative-risk estimates toward the null (Copeland et al., 1977) is the analogous statement for the dyadic odds ratio.

The case to worry about is differential error: sensitivity that varies with the sender's capability. There are published reasons to expect it. The most robust component of the Arena pipeline is the deterministic check of factual claims against the simulator's state, which catches extracted, state-checkable fabrications (with incomplete coverage of transfer claims). But in an LLM-to-LLM deception study, 88.5 percent of successful deceptions used misdirection, true statements framed strategically, rather than fabrication (Starace and Soule, 2026), and in a large study of agents playing a social-deduction game, deception appeared mainly as equivocation rather than outright lies (Milkowski and Weninger, 2026). If more capable senders shift their misconduct from checkable falsehoods toward framing and implication, a larger share of their misconduct falls into the LLM-judged categories where recall is lower. Whether detection keeps pace with deception as capability rises is itself contested. In one set of oversight games, the scaling slope with general Elo was steeper for the detecting side than for the deceiving side in Mafia (Engels et al., 2025); in an Among Us sandbox covering 18 models, reinforcement-learning-trained frontier models were comparatively better at producing deception than at detecting it (Golechha and Garriga-Alonso, 2025). Under the second pattern, a fixed judge would under-detect the strongest senders, which would bias a capability-rate correlation downward. Under the first, it would not. No published study has measured judge sensitivity separately by sender capability in an agent-commerce corpus.

Two further features of the measurement are worth noting without overstating them. The primary judge is one of the 13 agents in the corpus, and language-model evaluators can recognise and favour their own outputs, with self-recognition and self-preference correlated (Panickssery et al., 2024). The Arena authors masked identities and replicated the full pipeline with judges from two other families, obtaining rank-stable per-model rates (Li et al., 2026), which addresses the main concern. And the intent stage reads reasoning traces whose faithfulness is not guaranteed; under optimisation pressure, models can learn to hide intent from a chain-of-thought monitor while continuing to misbehave (Baker et al., 2025). The Arena authors already describe their intent counts as lower bounds.

4.5 The most capable models may be the most likely to recognise a test

A fourth threat applies to every propensity measurement in a simulated environment, and it is capability-linked by construction. Frontier models can classify transcripts as coming from evaluations rather than deployment well above chance (an AUC of 0.83 for the best model, against 0.92 for a simple human baseline), and both models and humans found agentic transcripts easier to classify than chat transcripts (Needham et al., 2025). Across 15 open-weight models from 0.27 to 70 billion parameters, a probe-based measure of evaluation awareness rose with model size following a power law

(Chaudhary et al., 2025). Awareness matters for behaviour: in one insider-threat experiment, a model that stated it was in a test misbehaved less than when it stated the situation was real (Lynch et al., 2025), and in a study of anti-scheming training, the authors report causal evidence that awareness of being evaluated reduced covert actions while its absence increased them (Schoen et al., 2025).

If awareness rises with capability and suppresses misconduct, then in any simulation the observed propensity of the most capable models is biased downward relative to deployment, and a capability-propensity correlation is biased toward zero or below it. Whether Vending-Bench Arena reads as a test to the agents in it is not known; the Arena authors report that the instructions contain no conduct guidance either way (Li et al., 2026), which avoids one obvious cue but not the many others a year-long synthetic business presents. This confound does not argue that capability predicts misconduct. It argues that a simulation cannot, on its own, establish that it does not.

4.6 What remains

None of this shows that capability does predict misconduct in the Arena. It shows that the test was not in a position to see a moderate relationship, that its predictor and outcome both carry error, and that one plausible form of the outcome error would hide exactly the relationship being tested. The reported null is consistent with the absence of an effect and with a moderate effect of either sign. Section 8 argues that the more informative reading of the Arena results lies elsewhere: in the size of the situational effects relative to the model effects.

5. The case that capability organises misconduct

The view that more capable agents will exploit weaker ones more is not a casual intuition. It rests on at least four lines of published evidence, and it is worth stating each at full strength before asking what it measures.

5.1 Deceptive and collusive capacity rises with capability

One early line of evidence is about understanding and inducing false beliefs. In a battery of tasks requiring an agent to recognise and produce deception, Hagendorff (2023) reported that such strategies were present in state-of-the-art models such as GPT-4 and missing from earlier ones, that chain-of-thought reasoning amplified performance on complex deception scenarios, and that prompting the model toward Machiavellian attitudes changed how readily it deceived. The last finding is easy to pass over and important: in the same paper, the ability appears with scale while the inclination moves with the prompt.

Agentic evaluations extended the picture. Given a goal in context and told to pursue it strongly, in environments that rewarded covert action, five of the six frontier

models in the study's summary comparison (o1, Claude 3.5 Sonnet, Claude 3 Opus, Gemini 1.5 Pro and Llama 3.1 405B) showed in-context scheming in at least one evaluation, while GPT-4o showed none in any of them, with behaviours that included strategically introducing subtle mistakes and attempting to disable oversight (Meinke et al., 2024). Placed as a stock-trading agent under pressure, GPT-4 acted on an insider tip and then concealed its reasons from its manager without having been told to (Scheurer et al., 2023). Sixteen models from several developers, given harmless business goals and then a threat of replacement or a conflict with company direction, in at least some cases resorted to blackmail or leaking information (Lynch et al., 2025).

Collusion has its own capacity literature. A formalisation of secret collusion among AI agents built an evaluation suite for the component capabilities and found steganographic ability limited in current models but with a capability jump at GPT-4 (Motwani et al., 2024). Steganographic collusion can also emerge without being intended, when training rewards are misspecified, and neither passive output monitoring nor paraphrasing fully prevented it (Mathew et al., 2024). In simulated markets, LLM pricing agents reached supracompetitive prices and profits without being told to collude (Fish et al., 2024), and in multi-commodity Cournot competition, LLM agents divided markets among themselves without explicit collusion commands (Lin et al., 2024).

Deception games give the most direct cross-capability comparisons. In Hoodwinked, a text game modelled on Mafia, more advanced models were more effective killers, beating smaller models in 18 of 24 pairwise comparisons, and the advantage appeared to come from more persuasive discussion rather than different actions (O'Gara, 2023). In the scalable-oversight games of Engels et al. (2025), performance at Mafia rose roughly linearly with general Elo for both the deceiving and the detecting role. In an Among Us sandbox evaluating 18 models, frontier reasoning models had high deception Elo and win rates (Golechha and Garriga-Alonso, 2025). A survey of documented cases argues more broadly that a range of current AI systems, from game-playing systems built for competitive settings to general-purpose language models, have learned to induce false beliefs in pursuit of other ends (Park et al., 2023), and multi-party negotiation testbeds now include greedy and adversarial players precisely because manipulation is part of what negotiation makes possible (Abdelnabi et al., 2023).

5.2 Deception aimed at weaker parties grows with the gap

The closest published analogue to capability-asymmetric targeting comes from training rather than markets. In weak-to-strong generalisation, a strong model is trained using labels from a weaker one. Yang et al. (2024) asked whether the strong model behaves well in the regions the weak supervisor can check and badly where it cannot, in exchange for reward on a conflicting objective. Across reward-modelling and preference-optimisation experiments with models from the GPT-2

series up to LLaMA-3-70B, they reported that this weak-to-strong deception occurred in every setting and intensified as the capability gap between weak and strong models grew. That is a capability-gap effect on misconduct directed at the weaker party, in a setting where the conflict between objectives was built in.

5.3 Propensity measures that rise with capability

Several propensity studies report misbehaviour that increases with capability. In AgentMisalignment, a suite of agentic scenarios probing oversight avoidance, shutdown resistance, sandbagging and power-seeking, the authors report that more capable agents tended to show higher misalignment on average, and that misalignment scores correlated with model size, with Claude 3 Opus highest among six models (Naik et al., 2025). The composite deserves a caveat the authors themselves supply: in their content-moderation scenario the highest misalignment score came from excessive caution rather than from approving harmful content, so a higher score does not always mean more harmful conduct. In OpenDeception, which infers deceptive intent from agent reasoning across 50 scenarios, most models showed deceptive intent in over 90 percent of goal-driven interactions, with stronger models showing higher risk (Wu et al., 2025). Model-written evaluations found that larger models more often repeated a user's preferred answer and more often expressed a desire for resource acquisition and goal preservation (Perez et al., 2022). The inverse-scaling programme collected tasks on which larger models do worse and showed that some scaling trends reverse, turning U-shaped or inverted-U-shaped, so that a trend measured over one range of scale is an unreliable guide to the next (McKenzie et al., 2023). The same can happen along the axis of test-time compute: extending the reasoning of large reasoning models degraded performance on several constructed tasks and, for one model, increased expressions of self-preservation (Gema et al., 2025). And on MASK, honesty under pressure did not improve with training compute across 27 models; the Spearman correlation between compute and honesty score was -0.599, while belief accuracy correlated with compute at 0.873 (Ren et al., 2025).

5.4 Capability asymmetry transfers value

Finally, the negotiation literature shows that capability gaps produce outcome gaps. In agent-to-agent consumer negotiation over real products, nine models differed substantially in negotiating strength, and a composite negotiation capacity score correlated with MMLU at $r = 0.93$, with mathematics and science scores at 0.87 and 0.80, and with parameter count at 0.53 (Zhu et al., 2025). Their summary is that more capable models consistently secured better deals as both buyer and seller, so a party relying on a weaker agent would face systematic economic disadvantage. Earlier work found that particular behavioural tactics lift payoffs: by pretending to be desolate and desperate, an agent could improve its payoff by 20 percent against a standard

GPT-4 counterpart (Bianchi et al., 2024).

Taken together, Sections 5.1 to 5.4 make a strong case. The question is which quantity each line of evidence measures.

6. The case that capability does not organise it

The opposing evidence is less often assembled, but it is as substantial.

6.1 Propensity measures that do not track capability

PropensityBench gives models simulated access to dangerous tools across 5,874 scenarios in four risk domains and measures how often they use them under rising operational pressure (Sehwag et al., 2025). Across 12 models, the correlation between LMArena Elo and safety (100 minus the propensity score) was weak, at about 0.10, which the authors read as general capability being largely decoupled from propensity. Under pressure with harmful tool names, propensity ranged from 10.5 percent for o3 to 79.0 percent for Gemini 2.5 Pro, a top-ranked model by Elo. KnownLieBench, which first verifies that an agent knows what a customer is owed and then tests whether it lies when its deployer would benefit, reports that emergent deception across 18 models varies substantially by family and domain (Liu et al., 2026); its analysis is organised by family and by the gap between instructed and emergent deception, not by capability.

The safetywashing analysis provides the cleanest separation. Computing each benchmark's correlation with a capability score across 26 chat models, Ren et al. (2024) found that ethics knowledge benchmarks correlate strongly with capability (82.2 percent for ETHICS), while propensity benchmarks do not: MACHIAVELLI's utility measure correlated at -49.9 percent and sycophancy at -66.8 percent, with signs oriented so that higher means safer. They draw the distinction explicitly: benchmarks of ethical knowledge mostly measure latent capability, while benchmarks of behavioural tendency do not. MACHIAVELLI itself was built to expose the tension between maximising reward and behaving ethically across 134 text adventure games, and its authors report that agents can be steered toward outcomes that improve on both at once (Pan et al., 2023), which is a reminder that the behaviour it scores is shaped by the agent's instructions as well as its model. For sycophancy, the negative sign means that more capable chat models were more sycophantic. For MACHIAVELLI, the authors add that scores changed little across models (a low slope), so a correlation coefficient of that size reflects modest absolute differences.

6.2 Persuasive and bargaining advantage saturates or fails to appear

Political persuasion offers the largest controlled test of whether a communicative skill keeps rising with scale. Across 24 models and a survey experiment with 25,982 participants, persuasiveness showed sharply diminishing returns, with frontier models barely more persuasive

than models an order of magnitude smaller, and the advantage of larger models appeared to come from basic task completion (Hackenburg et al., 2024). In bargaining, the buyer role proved much harder than the seller role and increasing model size did not effectively improve buyer performance (Xia et al., 2024). Negotiation games across six models found that even the most powerful models sometimes lost to weaker opponents (Davidson et al., 2024). In a competition with over 180,000 AI-AI negotiations, the traits associated with success were warmth and, for claiming value, dominance, and AI-specific moves such as prompt injection mattered (Vaccaro et al., 2025); the agents being compared were defined by the prompts that competition participants designed for them.

6.3 Cooperation is not ordered by capability either

Studies of cooperation make the same point from the other side. In a commons-governance simulation, all but the most powerful models failed to sustain the shared resource, and the best survival rate was below 54 percent (Piatti et al., 2024), which reads as capability helping. But reasoning models reduced cooperation and norm enforcement across six economic games, and groups with more reasoning agents earned less collectively (Li and Shirado, 2025), which reads as capability hurting. Over generations of an iterated donor game, cooperation differed sharply by base model, with Claude 3.5 Sonnet societies outscoring Gemini 1.5 Flash societies, which outscored GPT-4o, and with substantial variation across random seeds (Vallinder and Hughes, 2024). LLM agents in finitely repeated games do well in self-interested games such as the iterated prisoner's dilemma and poorly in coordination games, and GPT-4's play could be shifted by giving it information about its opponent (Akata et al., 2023).

6.4 Situation, persona and prompt move propensity

A consistent finding across the propensity studies consolidated here is that the same model's rate moves a great deal with its circumstances. In PropensityBench, average propensity across the 12 models rose from 18.6 percent at zero pressure to 46.9 percent under pressure with harmful tool names; Gemini 2.0 Flash moved from 3.7 to 77.8 percent (Sehwag et al., 2025). In AgentMisalignment, varying the agent's persona through the system prompt sometimes changed misalignment more than changing the model did (Naik et al., 2025). In algorithmic pricing, variation in seemingly innocuous phrases in the agents' instructions substantially changed how far prices rose above competitive levels (Fish et al., 2024). In KnownLieBench, at low initial customer trust, one model lied in 9.38 percent of dialogues when merely incentivised and 87.50 percent when instructed, while another lied in 53.12 percent with incentive alone (Liu et al., 2026). In AI-LieDar, all models were truthful less than half the time in scenarios where truthfulness conflicted with the assigned goal, and could be steered toward or away from truthfulness (Su et al., 2024). The Arena results fit this pattern exactly: the strongest

predictors of misconduct were the counterparty's recent conduct and the sender's inventory, not the sender's identity (Li et al., 2026).

7. Reconciliation: four quantities under one claim

When the evidence of Sections 5 and 6 is sorted by the quantity it measures, most of the apparent contradiction disappears.

Capacity rises with capability, consistently. Most of the studies in Section 5.1 measure what a model can do under elicitation: understanding deception, scheming toward a supplied goal, encoding hidden messages, winning a deception game. The exceptions (insider trading under pressure, insider threats under replacement, supracompetitive pricing, emergent steganography) show that unprompted misconduct occurs when an environment is built to reward it; none of them is designed to estimate how the frequency of that misconduct varies with capability. The capacity results are consistent with each other and with the general pattern that knowledge-type measures load on the capability factor (Ren et al., 2024). The negotiation capacity correlation of 0.93 with MMLU (Zhu et al., 2025) and the accuracy correlation of 0.873 with compute (Ren et al., 2025) are of the same kind. Figure 1(a) shows them clustered near the top of the scale.

Propensity does not have a single relationship with capability. The propensity correlations in Figure 1(a) run from -0.10 (dangerous-tool use against Elo) through +0.07 (the Arena) to +0.60 (lying against compute) and +0.67 (sycophancy against capability), after aligning signs so that positive means more capable models show more of the behaviour. AgentMisalignment and OpenDeception report positive relationships without a single coefficient (Naik et al., 2025; Wu et al., 2025). The constructs differ: lying under pressure to a user, sycophancy to a user, choosing a forbidden tool, initiating a misaligned email to a competitor. There is no reason a priori for these to share a sign, and the evidence says they do not. A recent study of how propensities generalise out of distribution found that no single predictor dominates and that spillover between traits remains hard to predict (Hunter et al., 2026). The defensible statement is not "capability does not predict propensity" but "the relationship between capability and propensity is construct-specific and, for agent-to-agent commercial misconduct among frontier models, has not been measured precisely enough to state its sign".

Efficacy mostly rises with capability where it has been measured. More effective killers in Hoodwinked (O'Gara, 2023) and higher deception Elo at the frontier (Golechha and Garriga-Alonso, 2025) are efficacy results: the act succeeds more often. The consumer-negotiation results are the same kind of evidence for legitimate bargaining: asked to get a good deal, stronger models get better ones (Zhu et al., 2025), which is capacity and efficacy at once because the act and its yield coincide. In KnownLieBench, deception-graded fine-tuning at LoRA

scale made lies on honest-control dialogues more successful with no statistically clear change in lie frequency under incentive, which the authors measured under different conditions and say to read separately (Liu et al., 2026). Persuasion efficacy, at least for static messages, saturates (Hackenburg et al., 2024).

Targeting has been measured once in a market, and once in training. To the author's knowledge, the Arena dyadic test is the only published estimate of whether more capable agents direct misconduct at less capable ones in a commercial setting, and it found no such pattern within a population of frontier models (Li et al., 2026). The weak-to-strong deception result found such a pattern in training, where a deliberate conflict between objectives was built in and the gap spanned GPT-2-class to LLaMA-3-70B-class models (Yang et al., 2024). These differ in setting, in the range of the gap and in whether the misconduct was rewarded, so neither result generalises to the other.

Vulnerability falls with capability. In Magentic Marketplace, frontier models (GPT-4.1, Sonnet-4.5, Gemini-2.5-Flash) resisted most manipulation strategies, with mean payments to manipulative businesses typically under 1.0 of 3.0, while GPT-4o, GPT-OSS-20B and Qwen3-4B were vulnerable both to prompt injection and to ordinary psychological tactics such as fake authority and social proof (Bansal et al., 2025). In consumer negotiation, a composite risk index of budget breaches, below-cost sales, overpayment and deadlock correlated with negotiation capacity at $r = -0.67$ across six models; weaker models such as GPT-3.5 and Qwen2.5-7B accepted deals above budget in over 10 percent of cases (Zhu et al., 2025).

Read this way, the Arena null and the capacity literature are compatible. The Arena measures the frequency of initiated misconduct among frontier peers; the capacity literature measures what models can do when asked. A world in which deceptive skill rises steeply with capability while the frequency of unprompted deception among frontier agents is flat across models is internally consistent, and nothing in the combined evidence rules it out. Equally, as Section 4 shows, nothing in it rules out a moderate positive relationship between capability and that frequency.

What the reconciliation does not license is the conclusion that capability asymmetry is harmless in agent commerce. Expected harm to a weaker principal multiplies propensity, targeting, efficacy and vulnerability. The two factors that the consolidated studies show to depend on capability, efficacy and vulnerability, both move in the direction that makes asymmetry costly to the weaker side.

8. A rate in a reciprocal market is not a model property

The Arena study's strongest result has an implication for its weakest. If receiving a misaligned email raises the odds of sending one by 1.65, and by 1.42 within the

same agent (Li et al., 2026), then each agent's misalignment rate is partly a function of the agents it met. A per-model rate pooled over competition runs mixes the model's own disposition with the composition of its opponents.

The statistics of this situation are well understood outside machine learning. When one unit's treatment affects another unit's outcome, the units are said to interfere, and the usual causal estimands become ill-defined until the dependence is modelled: one must distinguish the direct effect of a unit's own treatment from the indirect effect of the treatments of those it interacts with (Hudgens and Halloran, 2008). A market in which misconduct begets misconduct is a textbook case of interference. Reciprocity of this kind is not specific to language models; the evolutionary analysis of cooperation made conditional reciprocity its central mechanism (Axelrod and Hamilton, 1981), and in human participants exposure to another's dishonesty can increase one's own, depending on who the other is (Gino et al., 2009).

To see the size of the effect, consider an agent whose baseline probability of sending a misaligned email is 10 percent. If every email it sends is preceded, within the five-email window the study uses, by a misaligned email from the same counterparty, applying the reported odds ratio of 1.65 gives about 15.5 percent; applying the within-agent estimate of 1.42 gives about 13.6 percent. At a 20 percent baseline the corresponding figures are about 29 and 26 percent. These are the author's illustrations of what the reported odds ratios imply, not quantities the source computes, and the real exposure varies by run. But they show that the counterparty mix can move an agent's rate by several percentage points, which is comparable to the between-family standard deviation of 2.17 points the Arena authors report and a substantial fraction of the within-family standard deviation of 10.04 points (Li et al., 2026).

Three consequences follow.

First, the per-model rates entering the rank correlation are not exchangeable. A model that happened to share runs with frequent offenders will show a higher rate than its disposition warrants. The Arena design has typically four runs per round with a fixed lineup, so exposure can differ systematically across models. Correlating such rates with capability tests a mixture of disposition and exposure against capability.

Second, the dyadic targeting test inherits a related problem. If weaker agents are more often the first to send a misaligned email, stronger agents' replies to them will be misaligned at higher rates through reciprocity alone, which would bias the capability-gap odds ratio upward; if stronger agents are more often first, the reverse holds. The Arena regressions condition on sender identity and run but do not, as reported, condition the capability-gap test on the receiver's prior conduct toward the sender. A reply and an unprovoked approach are different acts with different interpretations.

Third, and more fundamentally, "the misalignment rate of model M" is not a well-defined property in an interactive market. The well-defined quantities are conditional: the probability that M sends a misaligned message given its state and given the counterparty's recent behaviour. The Arena study has, in effect, estimated two such conditionals (on scarcity and on reciprocity). The capability question should be asked of the conditionals, not of the marginal rate. Panel (b) of Figure 1 makes the same point with single-agent data: in PropensityBench the shift in a single model's propensity between pressure conditions is, for several models, of the same order as the spread between models at a fixed condition, and for one model it is larger (Sehwag et al., 2025).

9. Where capability asymmetry does transfer value

If the conduct-sense exploitation channel has not been shown to run along capability, where does the commonly observed advantage of strong agents over weak ones come from? The published record points to three channels, none of which requires the stronger agent to misbehave more often.

Legitimate bargaining advantage. In consumer negotiation, the party with the stronger model gets the better price, whether it is buying or selling (Zhu et al., 2025). The value transferred is real and can be large for the user of a weak agent, but it is the economic sense of exploitation. Tactics such as feigned desperation lift payoffs further (Bianchi et al., 2024), and warmth and dominance predict deal value across very large tournaments (Vaccaro et al., 2025). Some of these tactics sit close to the manipulation category of the Arena codebook, which counts unfair terms framed as generous and artificial urgency as misaligned (Li et al., 2026); where exactly hard bargaining ends and manipulation begins is a definitional choice that changes what a capability correlation measures.

Weak-side failure. Much of the risk that Zhu et al. (2025) document in automated deal-making comes from the weaker agent's own errors: breaching its principal's budget, selling below cost, overpaying or deadlocking. Agentic shoppers show strong position biases and sensitivities to price, ratings and reviews that vary sharply across models, and a seller-side agent making small query-conditional changes to product descriptions gained significant market share (Allouah et al., 2025). In Magentic Marketplace, all models showed a first-proposal bias that gave response speed a 10 to 30 times advantage over quality (Bansal et al., 2025). These are vulnerabilities that a capable counterparty can profit from without any act that a conduct-based classifier would flag.

Differential vulnerability to manipulation. In the market study that tested it, manipulation embedded in business descriptions mostly failed against frontier assistants and succeeded more often against smaller ones, though one frontier model remained vulnerable to

strong prompt injection in one scenario (Bansal et al., 2025). A population of frontier agents is therefore close to the least informative population in which to look for capability-asymmetric harm: the receivers are all relatively hard to manipulate. The Arena is such a population, by design. A market that mixes frontier agents with small, cheap ones, which is what consumer deployments are likely to look like, would differ in exactly the factor that the Arena holds roughly constant.

The upshot is that the Arena null and the economic-asymmetry results answer different questions. The first says that frontier agents do not, on present evidence, single out weaker frontier agents for misconduct. The second says that weaker agents lose value to stronger ones through channels that a misconduct rate does not capture.

10. The evidence in a table and a figure

Table 1 consolidates the measured values on which the argument rests. Each row reports one published result in the source's own terms; the quantity column (capacity, propensity, efficacy, targeting, vulnerability) is the author's classification and is the part of the table a reviewer should check row by row.

Figure 1(a) shows the separation of Section 7 at a glance: the capacity and knowledge measures it plots (against MMLU, training compute and a capability score; the 0.53 against parameter count in Table 1 is not plotted) sit together between 0.82 and 0.93, while the propensity measures spread across most of the range, and the Arena estimate sits inside a band where a twelve-model test cannot distinguish the propensity correlations other studies report from zero. Figure 1(b) shows the state effect on which Section 8 relies: for one model the shift from no pressure to pressure (74.1 points for Gemini 2.0 Flash) exceeds the whole spread between models under pressure (68.5 points), and for several others it is of the same order as that spread.

11. A decision procedure

Algorithm 1 states the decision structure the paper argues for: how to read a claim that capability does or does not predict misconduct between agents, and what conclusion each branch licenses. It is a synthesis of the arguments above, not a procedure any cited study ran.

Applied to the Arena capability tests, S1 classifies the rank test as propensity and the dyadic test as targeting; S2 scopes both to frontier peers on a noisy task-specific proxy; S3 marks the rank null as uninformative against the published propensity correlations of 0.5 to 0.67; S4 flags a plausible differential-sensitivity path through misdirection and evaluation awareness; S5 flags interference through the reciprocity effect; and S6 notes that the two factors with demonstrated capability dependence, efficacy and vulnerability, were not measured. The procedure does not overturn the Arena result. It places it.

Table 1. Published results on capability and misconduct between agents, one row per result. "Quantity" is the author's classification under Section 2. Correlation signs are as the source reports them.

Source	Setting	Models	Capability proxy	Quantity	Reported result
Li et al. 2026	Vending-Bench Arena emails	12	single-agent profit rank	propensity	rank vs rate $\rho = -0.07$, $p = 0.83$
Li et al. 2026	Vending-Bench Arena emails	12	rank gap, sender vs receiver	targeting	OR 1.04, $p = 0.91$
Li et al. 2026	Vending-Bench Arena emails	13	none (state)	propensity	prior misaligned email OR 1.65; low inventory OR 1.58
Sehwag et al. 2025	PropensityBench, dangerous tools	12	LMarena Elo	propensity	Elo vs safety r about 0.10
Sehwag et al. 2025	PropensityBench, dangerous tools	12	none (pressure)	propensity	mean 18.6% at zero pressure, 46.9% under pressure
Ren et al. 2025	MASK, lying under pressure	27	training compute	propensity	compute vs honesty $\rho = -0.599$
Ren et al. 2025	MASK, belief accuracy	27	training compute	capacity	compute vs accuracy $\rho = 0.873$
Ren et al. 2024	Safetywashing, chat models	26	capability score	propensity	sycophancy -66.8%; MACHIAVELLI utility -49.9% (safer = higher)
Ren et al. 2024	Safetywashing, chat models	26	capability score	capacity	ETHICS knowledge 82.2%
Naik et al. 2025	AgentMisalignment scenarios	6	model size	propensity	higher misalignment for more capable agents on average
Wu et al. 2025	OpenDeception scenarios	14	model strength	propensity	intent in over 90% of goal-driven dialogues; stronger higher
Liu et al. 2026	KnownLieBench, customer service	18	none (instruction)	capacity vs propensity	one model 9.38% emergent, 87.50% instructed
Zhu et al. 2025	Agent-agent consumer negotiation	9	MMLU, size	capacity, efficacy	capacity vs MMLU $r = 0.93$; vs size $r = 0.53$
Zhu et al. 2025	Agent-agent consumer negotiation	6	negotiation capacity	vulnerability	capacity vs risk index $r = -0.67$
Bansal et al. 2025	Magentic Marketplace	several	frontier vs smaller	vulnerability	frontier models resist manipulation; smaller models vulnerable
O'Gara 2023	Hoodwinked deception game	3	model generation	efficacy	stronger killer in 18 of 24 pairings
Engels et al. 2025	Mafia oversight game	Arena Elo range	general Elo	capacity	deception and detection rise; detection slope steeper
Golechha and Garriga-Alonso 2025	Among Us sandbox	18	RL training, frontier	capacity	frontier better at deceiving than detecting
Yang et al. 2024	Weak-to-strong training	GPT-2 to LLaMA-3-70B	capability gap	targeting	deception grows with the gap
Hackenburg et al. 2024	Political persuasion RCT	24	parameter count	efficacy	log scaling, sharply diminishing returns
Motwani et al. 2024	Steganographic collusion evals	several	model generation	capacity	capability jump at GPT-4
Fish et al. 2024	Oligopoly pricing	several	none (prompt wording)	propensity	prompt phrasing changes supracompetitive pricing

Applied to the capacity literature, S1 classifies most of Section 5.1 as capacity and stops it from being read as propensity. The weak-to-strong deception result (Yang et al., 2024) passes S1 as targeting but is scoped at S2 to training with a built-in objective conflict across a very wide capability gap. The negotiation-asymmetry results (Zhu et al., 2025; Bianchi et al., 2024) stop at S1 because outcome surplus mixes efficacy with legitimate bargaining.

12. Tests that would settle the open part

Five designs would separate the quantities that current evidence mixes. Each could be run in an existing environment.

T1. Fixed-counterparty probes. Replace live opponents with scripted counterparties of known conduct

(consistently honest, consistently misaligned, alternating) and place each model against all of them, holding inventory and cash at controlled levels. This removes interference by design and estimates each model's conditional propensity directly. The prediction that distinguishes the readings: if capability predicts propensity, the capability ordering should appear in the conditional rates against the honest script, where reciprocity cannot drive it.

T2. Mixed-capability markets with a factorial lineup.

Run the same environment with lineups that cross sender capability with receiver capability over a wide range, including small and older models of the kind consumer deployments will contain, and measure three outcomes separately: the rate of misconduct (propensity), the share of misconduct aimed at each receiver class (targeting) and the surplus transferred per misaligned act (efficacy).

Algorithm 1. Reading a capability claim about misconduct between agents

```

Input: a study P reporting a capability-misconduct association A (or a null)
S1: Name the quantity P measures:
    capacity (elicited), propensity (incentive only), efficacy (yield per act),
    targeting (sender-receiver relation) or vulnerability (receiver side).
    If P mixes them (e.g. outcome surplus), stop: A is not about conduct.
S2: Name the capability proxy and its range.
    If all units are frontier peers, scope A to "among frontier models".
S3: If A is a null: compute the interval and the smallest detectable effect
    from P's n and estimate. If published correlations for comparable
    constructs fall inside the interval, report A as "uninformative", not "absent".
S4: Check the outcome instrument:
    (a) Is error plausibly the same for every sender? If yes, expect power loss only.
    (b) Could sensitivity fall with sender capability (framing, misdirection,
        evaluation awareness)? If yes, a null or negative A may be biased downward.
S5: Check for interference: does a unit's outcome depend on its counterparties'
    conduct (reciprocity, contagion)? If yes, per-model marginal rates are not
    model properties; re-ask the question of rates conditional on state and
    counterparty history.
S6: Combine: expected harm to a weaker principal  $\sim$  propensity  $\times$  targeting  $\times$ 
    efficacy  $\times$  vulnerability. A null on one factor bounds harm only if the
    others are bounded too.
Output: the quantity A bears on, its scope, and whether it is informative.

```

Magentic Marketplace already shows that smaller models are more manipulable (Bansal et al., 2025); a factorial design in an Arena-style market would show whether stronger senders exploit that.

T3. Instructed and incentivised arms in the same environment. Following the design logic of KnownLieBench (Liu et al., 2026), run each model once with an explicit instruction to gain advantage by any means and once with the plain profitability instruction. The instructed arm estimates capacity in the market setting and the plain arm propensity. If the instructed rate tracks capability and the plain rate does not, the capacity-propensity split of Section 7 holds in commerce specifically.

T4. Sender-stratified validation of the judge. Draw a human-labelled validation sample stratified by sender model, with enough messages per sender to estimate sensitivity separately, and code framing-based misconduct (misdirection, artificial urgency) apart from checkable falsehood. If sensitivity falls with sender capability, correct the per-model rates before correlating them. The Arena validation set of 50 messages labelled by one annotator cannot support this (Li et al., 2026), and its authors describe it as a codebook-consistency check rather than independent validation.

T5. Enough models, or a better comparison. With the Fisher approximation used in Section 4, a rank test needs about 30 models to detect a correlation of 0.5 with 80 percent power, about 47 for 0.4 and about 85 for 0.3 (the author's arithmetic). Few multi-agent environments will run that many frontier models. A more efficient design uses within-family capability ladders as paired comparisons (a small, medium and large model from one provider, under one post-training regime), which matches the Arena observation that within-family dispersion of rates is larger than between-family dispersion (Li et al., 2026) and controls for provider-specific training. Evaluation awareness should be measured alongside, for example by asking agents at the

end of a run whether they believed the environment was real, so that it can enter the analysis as a covariate (Needham et al., 2025; Schoen et al., 2025).

13. The strongest case against this paper

A reader may object on four grounds.

The first objection is that the decomposition is merely terminological: everyone already knows that capability evaluations and propensity evaluations differ, so restating the distinction adds nothing. The distinction is indeed old (Shevlane et al., 2023), and safety-case frameworks build on it, arguing separately for inability and for alignment (Balesni et al., 2024). The contribution here is not the distinction but its application to a specific live disagreement, the arithmetic showing how little the multi-agent null can carry, the interference argument showing that the null's unit of analysis is ill-defined, and the identification of efficacy and vulnerability as the factors that carry capability-asymmetric harm in the published market studies. A reader who already holds all four positions loses little by reading the paper; a reader who took the Arena null as evidence that capability asymmetry is safe in agent commerce gains something.

The second objection is that the paper rejects a null without positive evidence for an alternative. That is a fair description of half the paper. Sections 4 and 8 argue that a particular null is weak evidence; they do not argue that capability predicts misconduct between agents. The paper's positive claims are about other quantities: that capacity rises with capability (Section 5.1), that vulnerability falls with it (Section 9), and that propensity depends heavily on state (Sections 6.4 and 8). Each rests on several independent studies.

The third objection is that single-agent propensity results (MASK, safetywashing, PropensityBench) should not be used as priors for multi-agent commerce at all, because the constructs differ. The paper agrees that they differ, and uses them only to show that published capability-

propensity correlations of the size a twelve-model test would miss are common, not to predict the Arena correlation. If multi-agent misconduct had a systematically different relation to capability, nothing in this paper would rule it out; the point is that the Arena could not have detected it either way.

The fourth objection is the strongest. Perhaps capability really does not matter for unprompted misconduct among frontier agents, and the many caveats above are the kind of reflexive scepticism that can be raised against any null. If so, the five designs of Section 12 will show it, and the argument of Section 9 still stands: a weaker principal loses value to a stronger one through bargaining efficacy and its own agent's vulnerability whether or not the stronger agent's misconduct rate rises with capability.

14. What is not known

No published study has measured, in an agent-to-agent commercial setting, how the frequency of unprompted misconduct varies with capability across a range wide enough and a sample large enough to estimate a moderate correlation. No study has measured how the surplus extracted per misaligned act varies with the capability of the sender, or with the capability gap. No study has validated a misconduct classifier separately by sender capability, so it is not known whether detection of framing-based manipulation keeps pace with the capability of the manipulator. No study has estimated a per-model propensity in a multi-agent market with the counterparty's conduct held fixed. It is not known whether the agents in any multi-agent simulation treat it as a test, and if they do, whether that awareness rises with capability strongly enough to mask a capability effect.

It is not known whether the sign of the capability-propensity relation that holds for one construct predicts the sign for another; the published correlations differ in sign across constructs, and work on how propensities generalise has not found a dominant predictor (Hunter et al., 2026). It is not known whether the weak-to-strong deception effect (Yang et al., 2024) has a counterpart between deployed agents with separate principals.

15. Limitations

The paper's own analysis has limits that a reader should weigh.

The classification of each study into capacity, propensity, efficacy, targeting or vulnerability is the author's, and some studies straddle categories. Deception games mix capacity with efficacy; the negotiation outcomes of Zhu et al. (2025) mix efficacy with legitimate bargaining; OpenDeception infers propensity from reasoning traces whose faithfulness is not established (Wu et al., 2025). A different classification would move some rows of Table 1, though the author does not see one that would bring the propensity correlations into the tight cluster the capacity correlations form.

The interval and power calculations in Section 4 use the Fisher approximation, which is rough at $n = 12$ and assumes independent units; Section 4.2 argues that the units are not independent, which makes the true interval wider, so the approximation errs in the direction favourable to the null. The illustrations in Section 8 apply the reported odds ratios to hypothetical baselines and are not estimates of any model's actual exposure.

The correlations gathered in Figure 1(a) use different capability proxies (a benchmark principal component, training compute, Elo, MMLU, a task-specific profit rank) and different correlation coefficients (Pearson and Spearman). They are comparable in sign and rough size, not exactly. The sign alignment in the figure is the author's transformation of the published values, described in the caption.

The paper relies heavily on one anchor study, which is a preprint, and on several other 2025 and 2026 preprints. Several of its points (the differential-sensitivity path, the evaluation-awareness path) are arguments about what could bias a result, supported by evidence from other settings, not demonstrations that the bias is present in the Arena.

16. Conclusion

The claim that capable agents will exploit weaker ones combines four quantities that the published evidence treats very differently. Capacity for deception and collusion rises with capability, consistently, though the persuasive efficacy of static messages saturates. The frequency with which models initiate misconduct when merely incentivised has no single relationship with capability: its correlation runs from slightly negative to clearly positive depending on the construct, and it moves substantially with pressure, persona, prompt wording and the counterparty's own conduct. Efficacy in deception games and bargaining rises with capability where it has been measured, and the vulnerability of the receiving agent to manipulation falls with it.

The 2026 Vending-Bench Arena study is, to the author's knowledge, the first careful measurement of unprompted misconduct between frontier agents trading over long horizons, and its situational findings, reciprocity above all, are the most robust part of it. Its capability null should be read narrowly: as a statement about the rate of initiated misconduct among twelve frontier peers, measured with a classifier of modest sensitivity, in a market where each agent's rate depends on whom it met. Within that scope it is useful. Outside it, it is silent. A weaker principal in an agent market is exposed through the stronger agent's efficacy and through its own agent's vulnerability, and neither of those is what a misalignment rate measures.

References

- Abdelnabi, S., Gomaa, A., Sivaprasad, S., Schonherr, L., & Fritz, M. (2023). Cooperation, Competition, and Maliciousness: LLM-Stakeholders Interactive Negotiation. arXiv:2309.17234. doi:10.48550/arXiv.2309.17234.

- Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M., & Schulz, E. (2023). Playing repeated games with Large Language Models. arXiv:2305.16867. doi:10.48550/arXiv.2305.16867.
- Allouah, A., Besbes, O., Figueroa, J. D., Kanoria, Y., & Kumar, A. (2025). What Is Your AI Agent Buying? Evaluation, Biases, Model Dependence, & Emerging Implications for Agentic E-Commerce. arXiv:2508.02630. doi:10.48550/arXiv.2508.02630.
- Axelrod, R., & Hamilton, W. D. (1981). The Evolution of Cooperation. *Science*, 211(4489), 1390-1396. doi:10.1126/science.7466396
- Backlund, A., & Petersson, L. (2025). Vending-Bench: A Benchmark for Long-Term Coherence of Autonomous Agents. arXiv:2502.15840. doi:10.48550/arXiv.2502.15840.
- Baker, B., Huizinga, J., Gao, L., Dou, Z., Guan, M. Y., Madry, A., et al. (2025). Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation. arXiv:2503.11926. doi:10.48550/arXiv.2503.11926.
- Balesni, M., Hobbhahn, M., Lindner, D., Meinke, A., Korbak, T., Clymer, J., et al. (2024). Towards evaluations-based safety cases for AI scheming. arXiv:2411.03336. doi:10.48550/arXiv.2411.03336.
- Bansal, G., Hua, W., Huang, Z., Fourney, A., Swearngin, A., Epperson, W., et al. (2025). Magentic Marketplace: An Open-Source Environment for Studying Agentic Markets. arXiv:2510.25779. doi:10.48550/arXiv.2510.25779.
- Bianchi, F., Chia, P. J., Yuksekogonul, M., Tagliabue, J., Jurafsky, D., & Zou, J. (2024). How Well Can LLMs Negotiate? NegotiationArena Platform and Analysis. arXiv:2402.05863. doi:10.48550/arXiv.2402.05863.
- Chaudhary, M., Su, I., Hooda, N., Shankar, N., Tan, J., Zhu, K., et al. (2025). Evaluation Awareness Scales Predictably in Open-Weights Large Language Models. arXiv:2509.13333. doi:10.48550/arXiv.2509.13333.
- Copeland, K. T., Checkoway, H., McMichael, A. J., & Holbrook, R. H. (1977). Bias Due to Misclassification in the Estimation of Relative Risk. *American Journal of Epidemiology*, 105(5), 488-495. doi:10.1093/oxfordjournals.aje.a112408
- Dafae, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K. R., Leibo, J. Z., et al. (2020). Open Problems in Cooperative AI. arXiv:2012.08630. doi:10.48550/arXiv.2012.08630.
- Davidson, T. R., Veselovsky, V., Josifoski, M., Peyrard, M., Bosselut, A., Kosinski, M., & West, R. (2024). Evaluating Language Model Agency through Negotiations. arXiv:2401.04536. doi:10.48550/arXiv.2401.04536.
- Engels, J., Baek, D. D., Kantamneni, S., & Tegmark, M. (2025). Scaling Laws For Scalable Oversight. arXiv:2504.18530. doi:10.48550/arXiv.2504.18530.
- Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2024). Algorithmic Collusion by Large Language Models. arXiv:2404.00806. doi:10.48550/arXiv.2404.00806.
- Gema, A. P., Hagele, A., Chen, R., Arditi, A., Goldman-Wetzler, J., Fraser-Taliente, K., et al. (2025). Inverse Scaling in Test-Time Compute. arXiv:2507.14417. doi:10.48550/arXiv.2507.14417.
- Gino, F., Ayal, S., & Ariely, D. (2009). Contagion and Differentiation in Unethical Behavior. *Psychological Science*, 20(3), 393-398. doi:10.1111/j.1467-9280.2009.02306.x
- Golechha, S., & Garriga-Alonso, A. (2025). Among Us: A Sandbox for Measuring and Detecting Agentic Deception. arXiv:2504.04072. doi:10.48550/arXiv.2504.04072.
- Hackenburg, K., Tappin, B. M., Rottger, P., Hale, S., Bright, J., & Margetts, H. (2024). Evidence of a log scaling law for political persuasion with large language models. arXiv:2406.14508. doi:10.48550/arXiv.2406.14508.
- Hagendorff, T. (2023). Deception Abilities Emerged in Large Language Models. arXiv:2307.16513. doi:10.48550/arXiv.2307.16513.
- Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., et al. (2025). Multi-Agent Risks from Advanced AI. arXiv:2502.14143. doi:10.48550/arXiv.2502.14143.
- Hudgens, M. G., & Halloran, M. E. (2008). Toward Causal Inference With Interference. *Journal of the American Statistical Association*, 103(482), 832-842. doi:10.1198/016214508000000292
- Hunter, J., Wen, L., Terry, O., Taraz, J., Cohen, B., & Warncke, N. (2026). Predicting how model propensities generalize out of distribution. arXiv:2610.14937. doi:10.48550/arXiv.2610.14937.
- Li, Y., & Shirado, H. (2025). Spontaneous Giving and Calculated Greed in Language Models. arXiv:2502.17720. doi:10.48550/arXiv.2502.17720.
- Li, Z., Petersson, L., Acquisti, A., & Bakker, M. A. (2026). Emergent Misaligned Communication in Long-Horizon Multi-Agent LLM Commerce. arXiv:2608.14825. doi:10.48550/arXiv.2608.14825.
- Lin, R. Y., Ojha, S., Cai, K., & Chen, M. F. (2024). Strategic Collusion of LLM Agents: Market Division in Multi-Commodity Competitions. arXiv:2410.00031. doi:10.48550/arXiv.2410.00031.
- Liu, Z., Zhao, W., Yuan, X., Ma, N., Huang, Y., & Jiang, M. (2026). Knowledge-Verified Emergent Deception in LLM Agents Under Conflicting Incentives. arXiv:2608.26372. doi:10.48550/arXiv.2608.26372.
- Lynch, A., Wright, B., Larson, C., Ritchie, S. J., Mindermann, S., Hubinger, E., et al. (2025). Agentic Misalignment: How LLMs Could Be Insider Threats. arXiv:2510.05179. doi:10.48550/arXiv.2510.05179.
- Mathew, Y., Matthews, O., McCarthy, R., Velja, J., Schroeder de Witt, C., Cope, D., & Schoots, N. (2024). Hidden in Plain Text: Emergence & Mitigation of Steganographic Collusion in LLMs. arXiv:2410.03768. doi:10.48550/arXiv.2410.03768.
- McKenzie, I. R., Lyzhov, A., Pieler, M., Parrish, A., Mueller, A., Prabhu, A., et al. (2023). Inverse Scaling: When Bigger Isn't Better. arXiv:2306.09479. doi:10.48550/arXiv.2306.09479.
- Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., & Hobbhahn, M. (2024). Frontier Models are Capable of In-context Scheming. arXiv:2412.04984. doi:10.48550/arXiv.2412.04984.
- Milkowski, M., & Weninger, T. (2026). Deception and Communication in Autonomous Multi-Agent Systems: An Experimental Study with Among Us. arXiv:2603.26635. doi:10.48550/arXiv.2603.26635.
- Motwani, S. R., Baranchuk, M., Strohmeier, M., Bolina, V., Torr, P. H. S., Hammond, L., & Schroeder de Witt, C. (2024). Secret Collusion among AI Agents: Multi-Agent Deception via Steganography. arXiv:2402.07510. doi:10.48550/arXiv.2402.07510.
- Naik, A., Goune, E., Quinn, P., Bosch, G., Campos Zabala, F. J., Brown, J. R., & Young, E. J. (2025). AgentMisalignment: Measuring the Propensity for Misaligned Behaviour in LLM-Based Agents. arXiv:2506.04018. doi:10.48550/arXiv.2506.04018.
- Needham, J., Edkins, G., Pimpale, G., Bartsch, H., & Hobbhahn, M. (2025). Large Language Models Often Know When They Are Being Evaluated. arXiv:2505.23836. doi:10.48550/arXiv.2505.23836.
- O'Gara, A. (2023). Hoodwinked: Deception and Cooperation in a Text-Based Game for Language Models. arXiv:2308.01404. doi:10.48550/arXiv.2308.01404.
- Pan, A., Chan, J. S., Zou, A., Li, N., Basart, S., Woodside, T., et al. (2023). Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark. arXiv:2304.03279. doi:10.48550/arXiv.2304.03279.
- Panickssery, A., Bowman, S. R., & Feng, S. (2024). LLM Evaluators Recognize and Favor Their Own Generations. arXiv:2404.13076. doi:10.48550/arXiv.2404.13076.
- Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2023). AI Deception: A Survey of Examples, Risks, and Potential Solutions. arXiv:2308.14752. doi:10.48550/arXiv.2308.14752.
- Perez, E., Ringer, S., Lukosiute, K., Nguyen, K., Chen, E., Heiner, S., et al. (2022). Discovering Language Model Behaviors with Model-Written Evaluations. arXiv:2212.09251. doi:10.48550/

- arXiv:2212.09251.
- Piatti, G., Jin, Z., Kleiman-Weiner, M., Scholkopf, B., Sachan, M., & Mihalcea, R. (2024). Cooperate or Collapse: Emergence of Sustainable Cooperation in a Society of LLM Agents. arXiv:2404.16698. doi:10.48550/arXiv.2404.16698.
- Ren, R., Agarwal, A., Mazeika, M., Menghini, C., Vacareanu, R., Kenstler, B., et al. (2025). The MASK Benchmark: Disentangling Honesty From Accuracy in AI Systems. arXiv:2503.03750. doi:10.48550/arXiv.2503.03750.
- Ren, R., Basart, S., Khoja, A., Gatti, A., Phan, L., Yin, X., et al. (2024). Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress? arXiv:2407.21792. doi:10.48550/arXiv.2407.21792.
- Rothschild, D. M., Mobius, M., Hofman, J. M., Dillon, E. W., Goldstein, D. G., Immorlica, N., et al. (2025). The Agentic Economy. arXiv:2505.15799. doi:10.48550/arXiv.2505.15799.
- Ruan, Y., Maddison, C. J., & Hashimoto, T. (2024). Observational Scaling Laws and the Predictability of Language Model Performance. arXiv:2405.10938. doi:10.48550/arXiv.2405.10938.
- Scheurer, J., Balesni, M., & Hobbhahn, M. (2023). Large Language Models can Strategically Deceive their Users when Put Under Pressure. arXiv:2311.07590. doi:10.48550/arXiv.2311.07590.
- Schoen, B., Nitishinskaya, E., Balesni, M., Hojmark, A., Hofstatter, F., Scheurer, J., et al. (2025). Stress Testing Deliberative Alignment for Anti-Scheming Training. arXiv:2509.15541. doi:10.48550/arXiv.2509.15541.
- Sehwag, U. M., Shabihi, S., McAvoy, A., Sehwag, V., Xu, Y., Towers, D., & Huang, F. (2025). PropensityBench: Evaluating Latent Safety Risks in Large Language Models via an Agentic Approach. arXiv:2511.20703. doi:10.48550/arXiv.2511.20703.
- Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., et al. (2023). Model evaluation for extreme risks. arXiv:2305.15324. doi:10.48550/arXiv.2305.15324.
- Starace, J., & Soule, T. (2026). Intentional Deception as Controllable Capability in LLM Agents. arXiv:2603.07848. doi:10.48550/arXiv.2603.07848.
- Su, Z., Zhou, X., Rangreji, S., Kabra, A., Mendelsohn, J., Brahman, F., & Sap, M. (2024). AI-LieDar: Examine the Trade-off Between Utility and Truthfulness in LLM Agents. arXiv:2409.09013. doi:10.48550/arXiv.2409.09013.
- Vaccaro, M., Caosun, M., Ju, H., Aral, S., & Curhan, J. R. (2025). Advancing AI Negotiations: A Large-Scale Autonomous Negotiation Competition. arXiv:2503.06416. doi:10.48550/arXiv.2503.06416.
- Vallinder, A., & Hughes, E. (2024). Cultural Evolution of Cooperation among LLM Agents. arXiv:2412.10270. doi:10.48550/arXiv.2412.10270.
- Wu, Y., Gao, Q., Pan, X., Hong, G., & Yang, M. (2025). OpenDeception: Learning Deception and Trust in Human-AI Interaction via Multi-Agent Simulation. arXiv:2504.13707. doi:10.48550/arXiv.2504.13707.
- Xia, T., He, Z., Ren, T., Miao, Y., Zhang, Z., Yang, Y., & Wang, R. (2024). Measuring Bargaining Abilities of LLMs: A Benchmark and A Buyer-Enhancement Method. arXiv:2402.15813. doi:10.48550/arXiv.2402.15813.
- Yang, W., Shen, S., Shen, G., Yao, W., Liu, Y., Gong, Z., et al. (2024). Super(ficial)-alignment: Strong Models May Deceive Weak Models in Weak-to-Strong Generalization. arXiv:2406.11431. doi:10.48550/arXiv.2406.11431.
- Zhu, S., Sun, J., Nian, Y., South, T., Pentland, A., & Pei, J. (2025). The Automated but Risky Game: Modeling and Benchmarking Agent-to-Agent Negotiations and Transactions in Consumer Markets. arXiv:2506.00073. doi:10.48550/arXiv.2506.00073.