Graded Expectations: Do Large Language Models Show Human-like Sensitivity to the Likelihood of Deceptive Speech Acts?

Xingyuan Zhao and Seana Coulson
University of California, San Diego
miozhao@ucsd.edu and scoulson@ucsd.edu

Abstract

Human discourse comprehension includes graded expectations about whether a speaker is likely to lie. If language models capture human-like discourse expectations, they should be sensitive not only to factual consistency but also to lie expectancy as a contextual probability from complex pragmatic cues. We test this idea using discourse scenarios with varying incentives to deceive. Human lie probability is estimated from free continuations, and model lie expectancy is derived from the probability mass assigned to human-produced lie versus truth continuations. Across Qwen3 models, likelihood-derived lie mass aligns strongly with human lie expectancy. The best performance comes from the base checkpoints. By contrast, post-trained and mode-specialized variants show weaker alignment. Qualitative analysis suggests a structured error pattern: models tend to overpredict lies when a response directly conflicts with known facts, but underpredict them when lie expectancy depends more on contextual pressures such as politeness, self-protection, or strategic gain. These results suggest that graded lie expectancy is recoverable from model continuation probabilities and can be learned, at least in part, through the ordinary next-token prediction objective.

1 Introduction

Human discourse comprehension is not limited to recovering the literal meaning of an utterance. In context, comprehenders also form expectations about why a speaker is saying something, whether the speaker is likely to be truthful, and how social or situational pressures shape the likely continuation of the interaction (Grice, 1975; Van Berkum, 2009; Hagoort et al., 2004). Lie expectancy is therefore not a peripheral phenomenon. It is part of ordinary discourse understanding: observers often anticipate that a speaker will lie in a given situation, even when they themselves know the truth, because they can infer the speaker's motives from the surrounding context.

This makes lie probability a meaningful target for language models. If a model captures human-like discourse expectations, it should be sensitive not only to factual consistency, but also to when deception is more or less likely in context. Recent work has shown that language models exhibit certain pragmatic abilities, including interpreting non-literal language and reasoning about speaker intent (Hu et al., 2023), yet their performance on tasks requiring inference about others' mental states remains mixed (Strachan et al., 2024; Kosinski, 2024; Sap et al., 2022; Ullman, 2023). Lies are shaped by structured pressures: avoiding punishment, protecting self-image, sparing another person's feelings, securing material gain, or manipulating what the listener will believe or do. These pressures depend on more than isolated lexical cues. They arise from the interaction of discourse facts, speaker goals, social relationships, institutional settings, and culturally familiar patterns of behavior. A model that is insensitive to this structure may detect contradiction, yet still fail to approximate human expectations about when a lie is likely.

At the same time, lie expectancy is not the same as binary lie detection. Determining that a particular response is false relative to a scenario is a simpler problem than estimating how likely a speaker is to lie in that scenario. The former can often be supported by recovering the relevant facts and detecting inconsistency. The latter is more demanding: it requires a graded expectation over possible continuations and depends on complex pragmatic factors. This distinction is central to the present study. Our aim is not merely to ask whether language models can identify a lie, but whether their continuation preferences recover the *graded human expectation* that a speaker will lie.

This framing motivates a likelihood-based approach. If lie expectancy is part of discourse pre-

diction, then it should be reflected in the probability distribution a model assigns to possible continuations. Rather than relying on explicit metalinguistic prompting—which has been shown to diverge from direct probability measurements, particularly for complex linguistic tasks (Hu and Levy, 2023)–we ask whether a model assigns greater probability mass to human-produced lie continuations in contexts where humans themselves are more likely to expect a lie. Under this view, deception expectancy is not a special-purpose judgment layered on top of language processing; it is part of the predictive structure of discourse itself.

To test this idea, we use discourse scenarios in which speakers have varying motivations to lie, and we estimate human lie probability from free completions collected for a subset of items. We then derive model lie expectancy from the probability mass assigned to lie-labeled versus truth-labeled human continuations. Because this measure is defined over continuation probabilities, it directly targets the model's latent predictive structure rather than its explicit verbal judgments.

We evaluate Qwen3 models (Qwen Team, 2025) across multiple scales and training conditions, including base checkpoints and post-trained checkpoints. Before the main alignment analysis, we also assess preliminary competence on lie–truth classification and discourse comprehension tasks, allowing us to distinguish competence of basic contextual understanding from competence to recover graded human lie expectancy.

The central question of the paper is therefore straightforward: *do language models capture human graded expectations about when a speaker will lie?* More specifically, we ask whether lie expectancy is recoverable from model continuation distributions, whether this alignment varies as a function of scale and post-training, and whether stronger general competence necessarily implies stronger alignment with human lie expectancy. Across the Qwen models evaluated here, we find that likelihood-derived lie mass aligns strongly with human lie expectancy, with the best performance coming from larger base checkpoints. This suggests that base models can align more closely with human behavior than their post-trained counterparts (Kim and Davis, 2025), and graded lie expectancy is, at least in part, a kind of discourse-level predictive structure that language models can acquire through next-token prediction over distributional regularities in context.

2 Dataset and Human Measures

2.1 Materials

The materials used here were adapted from an ongoing study investigating the online processing of lies and truths in socially biased discourse contexts. Each item consists of a short narrative scenario that establishes a social situation in which a speaker has a clear motivation to lie (e.g., to avoid punishment, to spare someone's feelings, or to appear favorably to others). The scenario is followed by a target sentence containing a critical word whose substitution determines whether the utterance is a lie or a truth relative to the scenario. Critical words occur in antonym pairs (e.g., *boring/interesting*), and the scenarios are likewise paired such that a critical word that produces a lie in one scenario produces a truth in the other, so that differences between conditions are not attributable to lexical properties of the critical words.

Example A:

> *While hanging out at the mall, Jake saw a beautiful leather jacket and decided to take it. He put it on and walked quickly out of the store. But, as Jake passed through the door, an alarm went off. A security guard stopped him and asked him where he got the jacket.*
>
> `Lie: Jake told the guard that he just bought the new jacket.`
>
> `Truth: Jake told the guard that he just took the new jacket.`

Example B:

> *Pete had just bought a brand new cellphone and was excited to show it to his friends. He knew they would be impressed if they thought he had taken it from the store without paying. When he showed it to them, they asked Pete where he got the cellphone.*
>
> `Lie: Pete told his friends that he just took the new cellphone.`
>
> `Truth: Pete told his friends that he just bought the new cellphone.`

The full stimulus set comprises 76 scenarios, each ending with one of two potential target sentences (a lie or a truth), and followed by one of the two possible probes (a plausible or an implausible

probe continuation). For example, following the lie in the sample stimulus in EXAMPLE A, the plausible continuation probe was "The security guard asked Jake to show him the receipt for the jacket," and the implausible continuation was "The security guard invited Jake to a slumber party." All 76 items are used in the preliminary competence evaluations (Experiment 1). We then collected and annotated free completions based on the scenario from 44 participants for a subset of 40 items, and we use this as the human lie expectancy in modeling the alignment between LLM lie mass and human lie expectancy(Experiment 2).

3 Language Model Selection

We evaluate models from the Qwen3 family (Qwen Team, 2025) in order to examine how alignment with human lie expectancy varies as a function of model scale and post-training regime.

We compare models across four scales (0.6B, 1.7B, 4B, and 8B) and across three training conditions: base, post-trained, and later checkpoint variants.

Base models. (*Qwen3-0.6B-Base*, *Qwen3-1.7B-Base*, *Qwen3-4B-Base*, *Qwen3-8B-Base*). Base models derive their continuation preferences from distributional semantics in the pre-training corpus, without the additional reshaping introduced by later post-training. Their next-token distributions should therefore provide the cleanest estimate of the model's underlying continuation preferences.

Post-trained models. (*Qwen3-0.6B*, *Qwen3-1.7B*, *Qwen3-4B*, *Qwen3-8B*). These models have undergone multi-stage post-training, including reasoning-oriented optimization, Thinking Mode Fusion, and general reinforcement learning. Thinking Mode Fusion integrates thinking and non-thinking behaviors within a single checkpoint. Such post-training may improve downstream capability while also reshaping token-level continuation probabilities. Comparing them against the base models therefore tests whether post-training preserves or distorts the continuation structure relevant to human lie expectancy.

4B checkpoint variants. (*Qwen3-4B-Thinking-2507*, *Qwen3-4B-Instruct-2507*). These checkpoints allow a more fine-grained comparison within the 4B line. Both are causal language models released after pretraining and post-training. Unlike the original post-trained Qwen3 models, `Qwen3-4B-Thinking-2507` supports only thinking mode, automatically inserts `<think>` in the default chat template, and is described as having increased thinking length for highly complex reasoning tasks. `Qwen3-4B-Instruct-2507`, in contrast, supports only non-thinking mode and does not generate `<think></think>` blocks. A dual-mode fused model may encode a continuation distribution that reflects a compromise between deliberative reasoning-style behavior and direct response behavior. A single-mode checkpoint, by contrast, should express a more specialized continuation distribution.

This sampling scheme distinguishes between two possibilities. Human-like lie expectancy may be best preserved in pretraining-induced continuation structure, in which case base models should align most closely with human behavior. Alternatively, post-training may sharpen the pragmatic and inferential structure relevant to deception, in which case post-trained or thinking-oriented variants should perform better. A further possibility is that post-training improves general capability while distorting the token-level continuation probabilities on which our measure depends.

4 Preliminary Competence Evaluation: Experiment 1

Before testing item-level alignment with human lie expectancy, we asked whether the evaluated models possess basic competence on two discourse comprehension tasks. The first is *lie/truth discrimination*: given a scenario and a candidate utterance, the model must determine whether the utterance constitutes a lie or a truth in that context. The second task is a *scenario coherence judgment*: given a preceding scenario and a probe sentence, the model must determine whether the probe is a plausible continuation of the story. While not the primary theoretical target of the paper, these tasks serve as diagnostic checks on the models' capacity to represent the facts presented in the scenarios. Accordingly, performance on these preliminary tasks will allow us to determine whether subsequent failures in the main alignment analyses are attributable to lack of contextual understanding.

4.1 Lie–truth discrimination

We first evaluated whether each model could distinguish lies from truths relative to the discourse scenario. For each item, the model was shown the

scenario together with a candidate target utterance and asked to classify the target as either a *lie* or a *truth*. This evaluation was conducted on the full 76-item stimulus set. Classification accuracy was initially computed across all items, and then recorded separately for lie and truth target sentences. We reasoned that a model that cannot reliably distinguish a lie from the truth would be unlikely to succeed on the graded lie-expectancy analyses in Experiment 2. The purpose of this discrimination task was thus diagnostic and served as a preliminary check on contextual understanding.

Qwen-family results are shown in Table 1 and differed markedly across the models. Several smaller or no-thinking configurations showed highly asymmetric behavior, often with inflated lie accuracy coupled with very poor truth accuracy. For example, in no-thinking mode, Qwen3-0.6B achieved chance-level overall accuracy (.500) with perfect lie accuracy and zero truth accuracy, indicating a strong bias to respond "lie." A similar pattern appeared in several base and no-thinking configurations. By contrast, thinking-enabled and larger variants performed far better than their base and no-thinking counterparts. The strongest overall performance came from Qwen3-8B in thinking mode (.882), followed by Qwen3-4B-Thinking-2507 (.848) and Qwen3-4B in thinking mode (.829). These results indicate that while many Qwen configurations possess basic contextual competence, this competence depends strongly on model scale and post-training regime.

4.2 Scenario coherence judgment

We next evaluated whether the models could distinguish contextually coherent from incoherent probe continuations. For each item, the model was given the scenario and a probe sentence and was asked to judge whether the probe was coherent with the discourse context. Because each scenario was paired with plausible and implausible probes in both lie and truth conditions, this task supplements the binary lie/truth task and provides a broader diagnostic of discourse-level comprehension. Accuracy was computed both overall and separately for plausible versus implausible probes within the lie and truth conditions, respectively.

Figure 1 shows that human comprehension performance remained substantially above all evaluated Qwen models. Human accuracy was near ceiling overall (All = .912), with high performance in both lie (.920) and truth (.904) conditions. Among Qwen models, performance varied by scale and mode: base variants were strongest in the 0.6B, 4B, and 8B families, whereas the 1.7B family showed a weaker and less consistent pattern. The best-performing models were Qwen3-4B-Base (.789), Qwen3-4B-Thinking-ckpt (.783), and Qwen3-8B-Base (.776), but all remained below the human benchmark. This confirms that even the strongest Qwen variants exhibit a substantial gap from human comprehension on the preliminary discourse task.

Taken together, the results of Experiment 1 suggest that a subset of the models possess nontrivial discourse competence on the materials used here. At the same time, the substantial variability across Qwen configurations indicates that model family, scale, and post-training regime materially affect even these prerequisite abilities. In Experiment 2 we go on to ask whether the representations that develop in these models align with *graded* human expectations that a given speaker will lie.

5 Does Model Lie Mass Align with Human Lie Expectancy? Experiment 2

Experiment 2 asks whether models' continuation probabilities align with graded human lie expectancy at the item level.

5.1 Human lie probability

For a subset of 40 items, we collected free completions from 44 participants. Each participant read the scenario and the target sentence prefix (up to but not including the critical word) and provided a free-form completion. The first author then coded each completion as a lie response, a truth response, or ambiguous/not applicable.

For each item i, human lie probability is defined as the proportion of lie responses among all unambiguous lie/truth responses:

$$\hat{p}_i^{\text{human}}(\text{lie}) = \frac{n_i^L}{n_i^L + n_i^T}, \tag{1}$$

where n_i^L and n_i^T denote the number of lie and truth responses for item i, excluding ambiguous/not applicable responses from the denominator. This quantity serves as the human target in the item-level alignment analyses reported in this experiment.

In addition, for each item we retain the full set of observed human completions $\mathcal{C}_i$, together with their lie/truth labels and multiplicity counts $m(c)$

Model	Condition	Accuracy	Acc_{Lie}	$\text{Acc}_{\text{Truth}}$
Qwen3-0.6B	instruct, no-thinking	0.500	1.000	0.000
Qwen3-0.6B	instruct, thinking	0.645	0.737	0.553
Qwen3-0.6B-Base	base	0.487	0.974	0.000
Qwen3-1.7B	instruct, no-thinking	0.493	0.947	0.039
Qwen3-1.7B	instruct, thinking	0.755	0.868	0.640
Qwen3-1.7B-Base	base	0.500	0.789	0.211
Qwen3-4B	instruct, no-thinking	0.539	0.789	0.289
Qwen3-4B	instruct, thinking	0.829	0.934	0.724
Qwen3-4B-Base	base	0.592	0.947	0.237
Qwen3-4B-Thinking-2507	thinking checkpoint	0.848	0.974	0.720
Qwen3-4B-Instruct-2507	instruct checkpoint	0.664	0.842	0.487
Qwen3-8B	instruct, no-thinking	0.612	0.750	0.474
Qwen3-8B	instruct, thinking	**0.882**	0.974	0.789
Qwen3-8B-Base	base	0.724	0.829	0.618

Table 1: Preliminary lie–truth discrimination results for Qwen-family models. "Condition" distinguishes decoding or model configuration conditions. Thinking-mode Qwen models generally outperform matched no-thinking and base variants, with the strongest overall performance from Qwen3-8B in thinking mode.

for each unique completion $c \in \mathcal{C}_i$. These completions serve as the candidate set for the model likelihood-based lie mass measure: rather than generating continuations freely, the model scores the probability of each human-produced completion, weighted by the number of participants who produced it.

5.2 Model Likelihood-Based Lie Mass

Rather than eliciting an explicit judgment from the model, we derive *model* lie expectancy from the probability mass assigned to human-produced lie versus truth continuations. For each item i, we construct a context as in (2).

$$x_i = \text{scenario}_i + \text{prefix}_i. \tag{2}$$

Let C_i be the set of human completions for item i. For each completion $c \in C_i$, we compute a length-normalized conditional log-likelihood:

$$s(c \mid x_i) = \frac{1}{|c|} \sum_t \log P(w_t \mid x_i, w_{<t}), \tag{3}$$

where $|c|$ is the number of tokens in completion c, and w_t is the t-th token of that completion. To convert these scores into relative weights over the observed completion set, we apply a count-weighted softmax:

$$\tilde{q}(c \mid x_i) = \frac{m(c) \exp(s(c \mid x_i))}{\sum_{c' \in C_i} m(c') \exp(s(c' \mid x_i))}, \tag{4}$$

where $m(c)$ is the number of human participants who produced completion c. We then define model lie probability as the total normalized mass assigned to lie-labeled completions:

$$\hat{p}_i^{\text{model}}(\text{lie}) = \sum_{c \in C_i^L} \tilde{q}(c \mid x_i), \tag{5}$$

where $C_i^L \subseteq C_i$ denotes the subset of lie-labeled completions. This measure is intended to capture implicit deception expectancy as encoded in the model's continuation preferences, rather than in its explicit verbal judgments.

We evaluate alignment between model-derived lie probability and human lie probability using four item-level metrics: Pearson correlation, Spearman correlation, mean absolute error (MAE), and root mean squared error (RMSE). Pearson correlation measures linear correspondence to human lie expectancy, whereas Spearman correlation measures rank-order agreement. MAE and RMSE quantify absolute deviation from the human values, with lower values indicating closer fit.

5.3 Item-Level Alignment Results

Table 2 reports alignment results for all Qwen-family models. Across models, likelihood-derived lie probability shows substantial positive correspondence with human lie expectancy, indicating that model continuation distributions recover graded structure in human deception expectations.

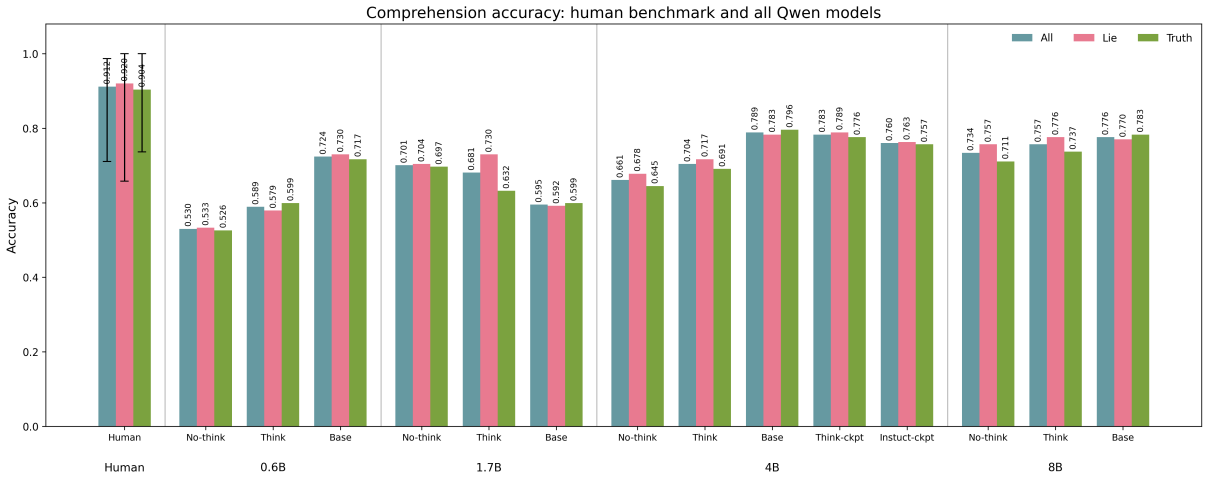

Figure 1: Human benchmark and Qwen-model performance on the preliminary comprehension task. Bars show overall accuracy (All), lie-condition accuracy (Lie), and truth-condition accuracy (Truth). Human values are means across participants, and error bars indicate the observed min–max range across participants with non-missing comprehension scores. Qwen models are grouped by parameter scale and mode.

Model	Pearson r	Spearman ρ	MAE	RMSE
4B-Base	**0.838**	**0.892**	0.166	0.233
8B-Base	0.806	0.852	**0.146**	**0.227**
0.6B-Base	0.788	0.829	0.173	0.260
8B	0.716	0.808	0.231	0.312
1.7B-Base	0.745	0.801	0.182	0.263
4B-Thinking-ckpt	0.728	0.797	0.213	0.299
0.6B	0.714	0.781	0.191	0.270
4B-Instruct-ckpt	0.690	0.780	0.231	0.325
4B	0.688	0.780	0.237	0.333
1.7B	0.611	0.720	0.250	0.360

Table 2: Likelihood-based alignment results for Qwen-family models on all items. Pearson and Spearman quantify item-level alignment with human lie probability; lower MAE and RMSE indicate closer absolute fit.

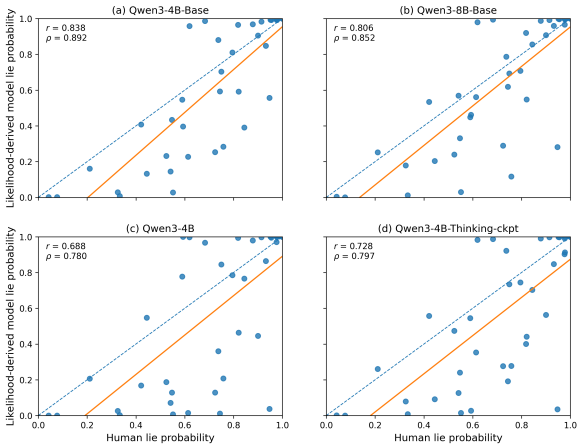

Figure 2: Item-level relationship between human lie probability and likelihood-derived model lie probability for representative Qwen models. The top row shows two base checkpoints (`Qwen3-4B-Base` and `Qwen3-8B-Base`), and the bottom row shows two non-base 4B variants (`Qwen3-4B` and `Qwen3-4B-Thinking-2507-checkpoint`). The dashed diagonal indicates perfect agreement ($y = x$), and the solid line shows the fitted linear trend.

The strongest performance is observed in the base checkpoints. In particular, `Qwen3-4B-Base` yields the strongest item-level alignment by correlation, achieving the highest Pearson r and Spearman ρ, whereas `Qwen3-8B-Base` yields the lowest absolute error, achieving the lowest MAE and RMSE.

More generally, base models show stronger alignment and lower absolute error than their post-trained counterparts at comparable scales. This pattern suggests that human-like lie expectancy is more faithfully preserved in pretraining-induced continuation structure than in distributions reshaped by later post-training.

Figure 2 illustrates the item-level relationship between model lie probability and human lie probability for representative models. Base checkpoints exhibit tighter alignment and less dispersion, whereas non-base variants show larger absolute deviations from the human values. Across panels, the fitted lines generally fall below the identity line, indicating systematic underestimation: the models track which items are more lie-favoring, but assign less extreme probabilities than humans do.

5.4 Qualitative analysis of high-disagreement items

We then took a closer look at the shared outliers among the best aligned Qwen models. The two

strongest shared overestimation items involve locally explicit contradictions between the discourse context and the speaker's response, and in both cases the models assign near-ceiling lie probability. In `overestimation items 1`, Brenda forgets to lock the door before a burglary. Her parents asked her if she had secured the door before leaving home. In `overestimation items 2`, Nancy has an eating disorder. She has eaten only an apple. Her doctor asks her what she ate that day. Why do the Qwen models expect lies in these scenarios when the humans do not? One possible explanation is that the models overweight the inferential structure in the scenario without fully capturing contextual pressures toward honesty(Sap et al., 2022; Ullman, 2023). The Brenda scenario, for example, includes explicit mention of her failure to lock the door followed by her parents query about this specific possibility. Moral norms surrounding interactions with one's parents may influence humans to entertain the possibility that Brenda might confess. By contrast, the models treat the conflict between the unlocked door implied by the first sentence and the locked door presupposed in the second as strongly predictive of a deceptive speech act. Likewise, whereas Nancy's eating disorder suggests she is likely to lie about her diet, the clinical interaction with her doctor may carry a competing norm of truthful disclosure. On this view, the models overestimate lie probability when local contradiction is strong but the social or institutional context leave room for honesty.

The shared underestimation items show the opposite pattern as the models underweight the social motivations to lie in the absence of conflicting factual information. In `underestimation items 1`, Rachel buys a painting at a low price. After it is damaged, the insurance agent asks how much she paid for the painting. In `underestimation items 2`, Brittany privately dislikes her friend's wedding dress, but knows it is too late to return the gown. Given she doesn't want to upset her friend, her friend asked Brittany what she thought of the dress. In both cases, the models assign substantially lower lie probability than human participants do. Unlike the overestimation items, the deceptive responses here are signaled less by blunt contradiction with an explicitly stated fact than by a socially or strategically motivated choice of utterance. In `underestimation items 1`, the relevant pressure comes from financial self-interest in an institutional setting; in `underestimation items 2`, it comes from face-saving politeness and the desire to avoid hurting another person's feelings. One possible interpretation is that the models are less sensitive to these higher-order pragmatic pressures than to direct factual conflict. On this view, the models underestimate lie probability when deception depends on social or strategic motivation rather than on an immediately salient contradiction with the discourse context.

Taken together, these cases suggest a common error pattern. The models are highly sensitive to explicit contradiction with known facts, but they underweight contextual factors that shape whether a speaker is expected to tell the truth or lie in a particular social setting. These interpretations remain tentative, but they suggest that model–human disagreement in the present task may arise in part from differences in how local factual conflict and higher-order pragmatic pressures are weighted.

6 Is Lie-Expectancy Alignment Reducible to General Competence?

A natural question is whether downstream alignment with human lie expectancy can be reduced to more general model competence. To examine this possibility, we conducted an exploratory cross-model comparison relating each Qwen model's Experiment 1 performance to its Experiment 2 alignment with human lie expectancy(Figure 3).

The two preliminary measures showed different patterns. Lie-classification accuracy was weakly negatively associated with downstream alignment, indicating that models that were better at binary lie–truth discrimination were not necessarily better at recovering graded human lie expectancy. By contrast, comprehension accuracy showed a moderate positive association with downstream alignment, suggesting that broader discourse-level understanding may be more relevant to human-like lie expectancy than categorical lie labeling alone.

These comparisons should be interpreted cautiously given the small number of models. Still, the pattern is informative. A plausible explanation is that binary lie–truth classification and lie-expectancy alignment rely on different levels of representation. Binary classification can be supported by recovering the relevant discourse facts in the scenario and determining whether a response is consistent with them. Lie-expectancy alignment is more demanding: it requires not only detecting factual inconsistency, but also modeling how likely

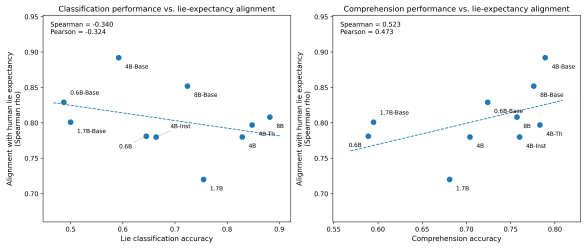

Figure 3: Exploratory cross-model relationship between preliminary task performance and downstream alignment with human lie expectancy in Qwen models. Left: lie-classification accuracy versus alignment with human lie expectancy. Right: comprehension accuracy versus alignment with human lie expectancy. Alignment is indexed by Spearman correlation between model-based lie mass and human lie probability. Dashed lines show least-squares fits for visualization only.

a speaker is to lie in that particular context. That, in turn, depends on pragmatic and discourse-level factors such as speaker goals, situational incentives, social pressures, and the relative plausibility of lie-favoring versus truth-favoring continuations. On this view, it is not surprising that downstream alignment is more closely related to comprehension accuracy than to binary lie-classification accuracy.

7 Conclusion

This paper asked whether language models capture graded human expectations about when a speaker is likely to lie in context. Across Qwen3 models, lie expectancy was recoverable from continuation probabilities over human-produced completions. The strongest alignment came from the base checkpoints, showing that stronger performance on post-trained or explicit task-style settings does not automatically yield more human-like lie expectancy.

The relevant signal depends on discourse-level context: model probabilities are conditioned on the scenario and utterance prefix rather than on isolated lexical material. The structured error pattern supports the same conclusion. Model errors were not random: the models were more sensitive to fact–response inconsistency than to contextual pressures such as politeness, accountability, or strategic self-interest. These findings are broadly compatible with psycholinguistic views in which comprehenders build structured discourse or situation-level representations that guide expectations about upcoming language (Zwaan and Radvansky, 1998; Kuperberg and Jaeger, 2016) .

The study also has clear limitations. The main results come from one model family, a relatively small item set, and a likelihood-based measure defined over observed human completions rather than unrestricted generation. Future work should test whether the same pattern holds across other architectures and should relate model errors more systematically to specific discourse variables, such as face-saving pressure, institutional incentives, and asymmetries in shared knowledge.

In sum, we find that likelihood-derived lie mass derived from the Qwen models aligns strongly with human lie expectancy, with the best performance coming from larger base checkpoints. This suggests that human-like lie expectancy is learnable, at least to some extent, from next-token prediction, not as an explicit theory of deception, but as a statistical regularity distributed across discourse, pragmatic, and situational structure.

References

H. Paul Grice. 1975. Logic and conversation. In *Speech Acts*, pages 41–58. Brill.

Peter Hagoort, Lea Hald, Marcel Bastiaansen, and Karl Magnus Petersson. 2004. Integration of word meaning and world knowledge in language comprehension. *Science*, 304(5669):438–441.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

Sanghee J. Kim and Forrest Davis. 2025. Discourse sensitivity in attraction effects: The interplay between language model size and training data. In *Proceedings of the Society for Computation in Linguistics*.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45).

Gina R Kuperberg and T Florian Jaeger. 2016. What do we mean by prediction in language comprehension? *Language, cognition and neuroscience*, 31(1):32–59.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? On the limits of social intelligence in large LMs. *arXiv preprint arXiv:2210.13312*.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriel Pansardi, Eugenio Scalber, Alessandro Ruber, Guido Manzi, Matilde De Luca, Gabriella Recchia, and Cristina Becchio. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Jos JA Van Berkum. 2009. The neuropragmatics of'simple'utterance comprehension: An erp review. In *Semantics and pragmatics: From experiment to theory*, pages 276–316. Palgrave Macmillan.

Rolf A. Zwaan and Gabriel A. Radvansky. 1998. Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2):162–185.