IPBench: Evaluating Text-Warranted Proposition Inference in Language Models

Anonymous authors
Paper under double-blind review

Abstract

Natural language inference (NLI) judges a hypothesis that is drawn from its premise and is true or false of a situation. We study the judgement of a hypothesis with the two opposite properties: it is exogenous, given before any premise is read, and it is abstract and ideational, a judgement that can be accepted or rejected but not verified. We call such a hypothesis an ideational proposition and the judgement of its relation to a text Text-Warranted Proposition Inference (TWPI): one of four labels, not applicable, accepted, rejected, or engaged without a determinate stance, which separates the prior question of engagement from the subsequent question of stance. To measure it we introduce IPBench: 516 cells formed from 172 passages of classical Chinese thought and 3 propositions, labelled by two experts under released semantic conventions ($\kappa = .92$), of which 174 are engaged, with a protocol that scores engagement and stance separately and measures over-engagement as a quantity of its own. Under one bare prompt frozen before any run, four models reach four-way accuracy of only .70 to .78, and the missing accuracy is over-engagement: every model engages more cells than the gold standard contains, by a factor of 1.03 to 1.33, while accept and reject are rarely confused, and two models route the surplus to the neutral label where two route it to a stance. Two preregistered controlled experiments find that prior knowledge, whether the model's memory of a passage's author or the orthodoxy of the proposition, has little and no consistent effect on the surplus. Applied over a fixed set of propositions, the judgement maps every text to a point of a shared proposition space in which the thought of different authors and periods can be compared; IPBench is not a new judge and its tables are not a ranking, but a testbed for the operation that such a space presupposes.

1 Introduction

Natural language inference (NLI) judges whether a hypothesis is entailed by, contradicts, or is neutral with respect to a premise, and on fact-level sentence pairs the task is mature (Bowman et al., 2015; Gururangan et al., 2018; McCoy et al., 2019; McKenna et al., 2023). Its hypothesis, however, is endogenous, drawn from or written against its premise, and it is a factual statement, true or false of a depicted situation. This paper studies the judgement of a hypothesis with the two opposite properties. (i) It is *exogenous*: fixed before any premise is read and put to every premise in the same words, so that whether the premise engages it at all is itself in question, a question NLI never asks. (ii) It is *abstract and ideational*: a judgement of value, priority or principle, stated in an analytic vocabulary that no text uses, which can be accepted or rejected but not verified, and which is reached by inference over the whole premise rather than by matching a sentence.

Borrowing NLI's frame, we call a hypothesis with these two properties an *ideational proposition* (IP), a judgement held rather than a fact reported (Chinese 思想命題), and we call the judgement of the relation between a premise text T and an IP P *Text-Warranted Proposition Inference* (TWPI). Its output is one of four labels, $\{\text{NA}, \text{ACCEPT}, \text{REJECT}, \text{NEUTRAL}\}$: not applicable, accepted, rejected, or engaged without a determinate stance. The formulation separates the prior question of whether T

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

engages P at all, *engagement*, which is where an exogenous hypothesis departs from NLI, from the subsequent question of which stance T takes, *stance*, which has the shape of NLI's three-way label (Section 2). Appendix A works through one passage of the corpus under both questions.

What evidence there is from the nearest tasks, models that label a pair as entailed whenever the hypothesis is familiar (McKenna et al., 2023) and that fail to abstain where abstention is the right answer (Kirichenko et al., 2025), suggests that language models are not equipped with this judgement (Appendix B). To measure it we introduce IPBench: 172 passages of classical Chinese thought $\times$ 3 IPs = 516 cells, 174 of them engaged, labelled by two experts under released semantic conventions ($\kappa = .92$; the first annotation is the gold standard, Section 3), with a protocol that scores engagement and stance separately and measures over-engagement, asserting a commitment the gold standard does not find, as a quantity of its own (Section 3).

Under one bare prompt frozen before any run, four models, `qwen3.8-max-0902`, `kimi-k3`, `deepseek-v4.1-flash` and `ZHIPU/GLM-5.3-FlashX`, reach four-way accuracy of only .70 to .78 on the 516 cells, and the missing accuracy is over-engagement (Section 4): 61% to 77% of each model's errors lie at the engagement boundary, every model engages more cells than the gold standard contains, by a factor of 1.03 to 1.33, and ACCEPT and REJECT are confused in at most nine cells per model. The surplus exits through different labels: `qwen3.8-max-0902` and `ZHIPU/GLM-5.3-FlashX` label it NEUTRAL (116 and 103 NEUTRAL predictions against 12 gold NEUTRAL cells), whereas `kimi-k3` and `deepseek-v4.1-flash` label it ACCEPT or REJECT. The controlled experiments of Section 5 indicate that prior knowledge, whether the model's memory of the passage's author (text level) or the orthodoxy of the proposition (proposition level), has little and no consistent effect on the surplus.

Beyond the single judgement, applying TWPI over a fixed set of IPs maps every text to a point of the *proposition space* (Section 2.3), a shared representation in which the thought of different authors, periods and civilisations can be represented and compared.

Our contributions are:

1. **An addition to NLI.** The ideational proposition, a hypothesis that is exogenous and abstract and ideational, and the judgement TWPI that relates a text to it in two steps, engagement and stance, with a positive engagement criterion given by the commitment positions Π_P (Section 2).

2. **A benchmark and a protocol.** IPBench, 516 cells labelled by one expert under released conventions, scored in two layers with over-engagement measured by OAI and FAS and a per-metric noise floor from independent reruns; under it four models over-engage, and the surplus exits through model-specific labels (Sections 3–5).

3. **A shared representation space.** The proposition space, in which the judgement gives each text coordinates comparable across authors and periods (Section 2.3).

2 Task Definition

2.1 Ideational Propositions

An IP can be looked at from two sides. From the text's side it is a hypothesis with two properties that the NLI hypothesis lacks: it is *exogenous*, and it is *abstract and ideational*. From the side of text comprehension it is a schema in the sense of Kintsch & van Dijk (1978), a reading goal that fixes what a text says about one question. The definition, fixed before annotation, reads:

Ideational Proposition (IP). A mid-level unit of ideational content lying between a concrete textual claim and a highly general ideational construct. Through controlled abstraction over concrete historical expressions it represents one substantive judgement, about the world, persons, social relations, value priorities, causal mechanisms, institutional arrangements or principles of action, possessing a determinate relational structure and capable of being affirmed or rejected. An IP must be abstract enough that heterogeneous expressions by different authors, texts, periods and cultures become comparable along a single ideational dimension, and contentful and historically traceable enough to distinguish substantively different beliefs, norms and inferential structures. For a text T, the judgement of Section 2.2 applied to every IP yields its *ideational summary* $\delta(T)$: the stance T takes on each IP it engages, NA on the rest (Section 2.3). IPs are the units from which every $\delta(T)$ is composed, and they are the stable nodes of the human thought space.

Two properties, each against a neighbour. Against a *topic* or a *key point*, an IP is *exogenous*: a topic is a word the text uses and a key point is induced from the corpus it summarises, whereas an IP is written before any text is read and put to every text in the same words, so the first question about any text is whether the IP is engaged at all. Against a *claim*, the unit of argument mining, an IP is *abstract and ideational*: a claim is a span of the text and is true or false of a situation, whereas an IP is a judgement of value, priority or principle, stated in an analytic vocabulary that no text uses, which can be affirmed or rejected but not verified and is therefore reached by inference over the whole passage. The nearest neighbour, the *latent intention* of Attributional NLI (Quan et al., 2026), is likewise inferred rather than stated but has neither property: its candidate set is drawn up per agent and per situation, and its object is a mental state of the speaker, a fact checked against a hidden role rather than a judgement to be affirmed or rejected. Att-NLI asks *why* the text was produced; TWPI asks *what* the text stands committed to (Table 3, Appendix B).

Each IP is a schema. Kintsch & van Dijk (1978) derive a text's *macrostructure* from the propositions it states, the textbase, by macrorules controlled by a *schema* that formalises the reader's goal; the macrostructure is implied by the textbase, it is layered with a single proposition at the top, and if the controlling schema is not determinate, macroprocessing is indeterminate (pp. 366, 372–374). An IP supplies exactly such a schema, $\sigma_P =$ "what does this text say about P": under it the propositions that bear on P survive compression, the rest are deleted, and the macrostructure of T is either empty or ends in one proposition about P. One step is ours: we fix the vocabulary of the top layer in advance to P's commitment positions Π_P (Section 2.2), so that a non-empty top layer is one of those positions and $F(T, P)$ reads it off (Figure 1a; the four steps are spelled out in Appendix C).

2.2 The Judgement

NLI's hypothesis is a statement written against the premise; TWPI's hypothesis is an exogenous IP that no sentence states, so its inference runs from a whole passage to a node given in advance, and whether the node is present at all is itself in question. For a text T and a proposition P the annotator (or system) returns

$$F(T, P) \longrightarrow \{\text{NA},\ \text{ACCEPT},\ \text{REJECT},\ \text{NEUTRAL}\}, \tag{1}$$

where NA means the text does not engage the proposition, ACCEPT and REJECT that it commits for or against it, and NEUTRAL that it engages the proposition without taking a determinate side.

Two steps: engagement, then stance. *Engagement* (NA vs. the rest) answers to the exogenous property and needs a positive criterion, because NEUTRAL defined as the residue between NA and a determinate side fails in practice (Section 5). Every proposition determines a small set of *commitment positions* Π_P, the distinct stances a text can take up towards P's relational structure, such as endorsing it, denying it, reversing its ordering, or treating its object as bearing on the judgement without ranking it; T engages P if and only if T undertakes a commitment that occupies one of these positions, and NA is the case in

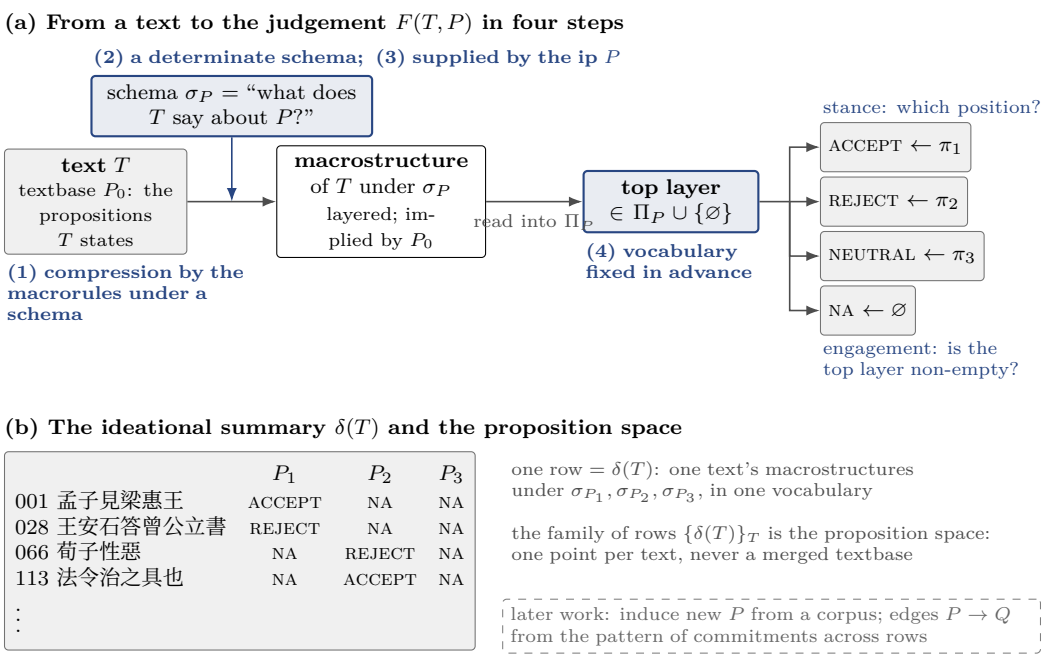

Figure 1: From a text to the judgement, and from the judgement to the proposition space. (a) The four steps of Section 2.1: the textbase of T is compressed by the macrorules under a schema (1); the schema must be determinate (2) and the IP supplies it as σ_P (3); the vocabulary of the top layer is fixed in advance to the commitment positions Π_P (4), so that engagement asks whether the top layer is non-empty and stance asks which position it holds. (b) One row of the matrix is a text's ideational summary $\delta(T)$; the family of rows is the proposition space of Section 2.3. The rows shown are gold labels of four passages of IPBench. Inducing new propositions and drawing edges between them are later work.

which it occupies none, which includes mentioning P's object without committing to anything about it. *Stance* (ACCEPT / REJECT / NEUTRAL, conditional on engagement) answers to the abstract and ideational property and records which position the commitment occupies: ACCEPT and REJECT are the positions for or against P as a whole, NEUTRAL any other occupied position. The two steps are different in kind, Section 3 scores them separately, and Section 4 shows that most of each model's errors fall on the first. The quantity underlying engagement is a degree of commitment thresholded into a label, which is why a difference between settings is read as a move of the operating point until shown otherwise. The position lists of this release are given, with their provenance, in Appendix C.

2.3 Proposition Space

Let $\mathcal{P}$ be the set of IPs. The *ideational summary* of a text is the function $\delta(T)\colon \mathcal{P} \to \{\text{NA}, \text{ACCEPT}, \text{REJECT}, \text{NEUTRAL}\}$, $\delta(T)(P) = F(T, P)$, one judgement per proposition; in the terms of Section 2.1, $\delta(T)(P)$ is the top layer of the macrostructure of T under σ_P, read into Π_P. Because $\mathcal{P}$ is exogenous, every text is judged on the same axes, and texts from different authors, periods and civilisations become points of the same space $\{\text{NA}, \text{ACCEPT}, \text{REJECT}, \text{NEUTRAL}\}^{\mathcal{P}}$, which we call the *proposition space*. The proposition space of a corpus is the family $\{\delta(T)\}_T$ indexed by text, a text-by-proposition matrix (Figure 1b), and never a merged textbase: the macrorules presuppose one discourse, and passages by different authors share neither an order nor a coherence graph (Kintsch & van Dijk, 1978, pp. 365–368). Writing P_0 for the propositions T states, K for the reader's knowledge, $\text{Cn}(\cdot)$ for logical closure and $W(T; \sigma_P, K)$ for the content T warrants under σ_P,

$$P_0 \;\subseteq\; W(T; \sigma_P, K) \;\subseteq\; \text{Cn}(P_0 \cup K), \tag{2}$$

and the one respect in which this differs from van Dijk's construction carries the whole difficulty: a macroproposition is *derived* from the text, whereas an IP is *given*, and the judgement asks whether the text's warranted content reaches it. NA covers a candidate outside the closure as well as one inside the closure but outside W, and over-engagement places in W a candidate that belongs outside it (Appendix D). Inducing new propositions from a corpus and drawing edges between them from the pattern of commitments are later work.

3 Benchmark and Metrics

3.1 Corpus and Propositions

Corpus. The source corpus is 172 passages of classical Chinese thought drawn from the Confucian volumes of the *Zhonghua Dadian, Philosophy Canon* (《中華大典·哲學典·儒家分典》), ranging from 36 to 2,208 characters (median 313.5, mean 403.2). Each passage is released as one text with a numeric identifier, and every passage is paired with every proposition. The corpus was assembled as candidate-relevant passages for the three propositions (the selection record of the candidate pool was not kept, so the funnel cannot be reported), so the engagement rate here ($174/516 = 33.7\%$) is far above that of an unselected corpus, and every precision figure in this paper is to be read against this base rate. The passages and their commentaries are almost certainly in the pretraining data of every model evaluated; what is new is the pairing with an exogenous proposition and the label, and Section 5 tests directly whether a model's memory of a passage's source drives its errors.

Propositions. The 3 propositions, with the Chinese wording used in annotation and in the evaluation prompts (the template in Appendix J words P_1 differently):

P_1 The moral obligations that rulers, governments or elites owe to society are more important than their economic obligations (君主、政府或精英对社会负有的道德义务比经济义务更重要).

P_2 Social order ought to depend on morality more than on law (社会秩序对道德的依赖应该大于法律).

P_3 Artisans and technical skill deserve respect (工匠与技艺应当得到尊重).

P_1 and P_2 are comparative; P_3 is a single-argument evaluative predicate. Three semantic conventions fixed in the written protocol resolve recurring interpretive questions, among them that 「利」(*li*, benefit or profit) counts as covering economic obligation; they are released with the data and are part of what a system is asked to apply (Appendix G).

3.2 Annotation

Protocol and gold standard. For each passage and each proposition the protocol asks for one label from the closed set $\{\text{NA}, \text{ACCEPT}, \text{REJECT}, \text{NEUTRAL}\}$ (不适用、认可、拒斥、无明确立场), reached in five stages: semantic normalisation of the proposition; the engagement decision, whether the passage engages (激活, "activates") the proposition at all; the stance decision, conditional on engagement; verification; and calibration (Appendix G). Two experts, both with training in intellectual history and research experience, labelled every cell back to back, after reaching a shared understanding of the four labels and of the three conventions. Over the 516 cells they give the same label on 95.9% (Cohen's $\kappa = .920$ (Cohen, 1960); nominal Krippendorff's $\alpha = .920$). Split by step, the applicability κ, with the three engaged labels collapsed against NA, is .961, and the conditional stance κ on the cells both experts mark as applicable is .863; per proposition, the overall κ is .893, .922 and .948, the applicability κ .963, .962 and .957, and the conditional stance κ .787, .860 and .959 for P_1, P_2 and P_3: the two experts' results are highly consistent at both steps and on every proposition. The first expert's annotation is the gold standard: every one of its 516 cells carries exactly one label from the closed set (0 violations), and every κ reported for a model in this paper is model–gold agreement.

NA is two thirds of the cells and NEUTRAL has 12 cells in all; the marginal distribution and its consequences for the protocol are in Appendix H (Table 4).

3.3 Evaluation Dimensions

The unit of evaluation is a (text × proposition) pair. Let $\mathcal{D}$ be the set of all cells and $N = |\mathcal{D}|$. Cell i has gold label y_i and prediction $\hat{y}_i$, with $y_i, \hat{y}_i \in \mathcal{Y} = \{\text{ACCEPT}, \text{REJECT}, \text{NEUTRAL}, \text{NA}\}$. The activation indicators are $a_i = \mathbf{1}[y_i \neq \text{NA}]$ and $\hat{a}_i = \mathbf{1}[\hat{y}_i \neq \text{NA}]$; the gold-activated cells are $\mathcal{D}^+ = \{i : a_i = 1\}$ and $\mathcal{D}^- = \mathcal{D} \setminus \mathcal{D}^+$.

Following the two-step decomposition of Section 2, evaluation has three dimensions (Table 5, Appendix I). The **four-way layer** compares y with $\hat{y}$ on $\mathcal{D}$; the **activation layer** compares a with $\hat{a}$ on $\mathcal{D}$; the **conditional stance layer** compares y with $\hat{y}$ on $\mathcal{D}^+$ only. Every dimension is reported both pooled (Overall) and per proposition (P_1–P_3); the per-proposition figures restrict $\mathcal{D}$ to that proposition's cells and apply the same formulas. Engagement is scored as the activation layer; we keep that name for the layer and for OAI and FAS in the tables, and write engagement, over-engagement and false engagement in the text.

3.4 Metrics

Generic metrics (four-way and conditional stance layers). Let $\mathcal{S}$ be the evaluation set ($\mathcal{D}$ or $\mathcal{D}^+$), $n = |\mathcal{S}|$, and $\mathcal{C}$ the set of classes that occur in the gold or in the predictions on $\mathcal{S}$; $\text{TP}_c, \text{FP}_c, \text{FN}_c$ are the one-versus-rest counts for class c, $n_c = |\{i : y_i = c\}|$ and $\hat{n}_c = |\{i : \hat{y}_i = c\}|$.

$$\text{Acc} = \frac{1}{n} \sum_{i \in \mathcal{S}} \mathbf{1}[\hat{y}_i = y_i], \tag{3}$$

$$\text{Macro-F1} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{2\,\text{TP}_c}{2\,\text{TP}_c + \text{FP}_c + \text{FN}_c}, \tag{4}$$

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \qquad p_o = \text{Acc}, \qquad p_e = \frac{1}{n^2} \sum_{c \in \mathcal{C}} n_c\, \hat{n}_c. \tag{5}$$

The conditional stance layer does not discard missed engagements: a cell $i \in \mathcal{D}^+$ with $\hat{y}_i = \text{NA}$ counts as an error, in which case NA enters $\mathcal{C}$ with an F1 of 0. κ is Cohen's κ (Cohen, 1960); the gold standard is one expert's annotation, so every κ reported for a model is model–gold agreement, distinct from the inter-annotator κ of Section 3.2.

Activation layer. With a as gold and $\hat{a}$ as prediction, the binary counts $\text{TP}, \text{FP}, \text{FN}, \text{TN}$ (positive class = activated) give

$$\text{Prec} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{Rec} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \qquad \text{F1} = \frac{2\,\text{Prec} \cdot \text{Rec}}{\text{Prec} + \text{Rec}}, \tag{6}$$

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}, \tag{7}$$

and κ as above, applied to $(a, \hat{a})$. Two further metrics target over-engagement:

$$\text{OAI} = \frac{\sum_{i \in \mathcal{D}} \hat{a}_i}{\sum_{i \in \mathcal{D}} a_i} = \frac{\text{TP} + \text{FP}}{\text{TP} + \text{FN}}, \tag{8}$$

$$\text{FAS} = \frac{|\{i \in \mathcal{D} : y_i = \text{NA},\ \hat{y}_i \neq \text{NA}\}|}{|\{i \in \mathcal{D} : \hat{y}_i \neq y_i\}|} = \frac{\text{FP}}{\text{number of four-way errors}}. \tag{9}$$

OAI is the ratio of predicted to gold engagements: 1 means calibrated in volume, > 1 over-engagement. FAS is the share of the four-way errors that are false engagements; it answers on which layer the error budget is spent.

Metrics independent of the operating point. OAI, precision and recall move with the operating point and are not evidence of capability on their own. A system that exposes a

continuous activation score s_i reports

$$\text{AUC} = \frac{1}{|\mathcal{D}^+|\,|\mathcal{D}^-|} \sum_{i \in \mathcal{D}^+} \sum_{j \in \mathcal{D}^-} \Big(\mathbf{1}[s_i > s_j] + \tfrac{1}{2}\,\mathbf{1}[s_i = s_j] \Big). \tag{10}$$

A system that outputs discrete labels only cannot compute it, and the tables mark it as not applicable.

4 Results

4.1 Setup

The baseline treats each model as a zero-shot annotator: one stateless call per (text × proposition) cell, one fixed template, a closed-set label as the only output, the standard way of using a language model in place of a human annotator (Gilardi et al., 2023; Zheng et al., 2023). The template (Appendix J) presents the passage and the proposition with no examples, defines the four labels through what the speaker undertakes (Brandom, 1994), and asks for one JSON object with the label first and a rationale second. Four models, `qwen3.8-max-0902`, `kimi-k3`, `deepseek-v4.1-flash` and `ZHIPU/GLM-5.3-FlashX`, are called through one OpenAI-compatible interface at temperature 0 and seed 0, with the same parsing rule and an invalid-cell rule fixed in advance; `glm-5.3-prime`, run under the same rules, appears as a fifth row of the tables. Each model has two settings that differ only in the thinking switch: thinking off is the main setting and thinking on the comparison setting, since reasoning fine-tuning is known to worsen abstention failures (Kirichenko et al., 2025); the GLM interface cannot disable thinking, so the two GLM models have a thinking setting only, which serves as their main setting. Each main setting was run a second time independently, and the largest absolute difference between the two runs, per pooled metric, is this paper's *noise floor* (Table 20, Appendix M): four-way accuracy .014, κ .031, precision .016, recall .034, OAI .040, FAS .032 and conditional stance accuracy .046; no smaller difference is reported as a finding. The readings that decide the findings below were written before the runs; the full protocol, the pre-registration record and the deviation log are in Appendices K and L.

4.2 Results

Evaluation follows Section 3: the four-way layer, the activation layer (with OAI and FAS) and the conditional stance layer. The baseline outputs discrete labels only, so no AUC is reported, and every Cohen's κ is model–gold agreement (Section 3). Table 1 gives the pooled metrics of every setting and Table 2 where the four-way errors of each main setting go; the stability table, the per-proposition tables, the confusion matrices and the further pre-registered readings are in Appendix M. The findings below are on the main setting of the four reference models, single run, every difference read against the noise floor; the pre-registered decision rules behind them are in Appendix L.

The errors lie on the activation layer. False engagements (gold NA judged engaged) and missed engagements (gold engaged judged NA) make up 61%, 77%, 71% and 68% of the four-way errors of `qwen3.8-max-0902`, `kimi-k3`, `deepseek-v4.1-flash` and `ZHIPU/GLM-5.3-FlashX` (Table 2; this order throughout), while ACCEPT and REJECT are confused in at most 9 cells in any model.

The direction is over-engagement. OAI is 1.230, 1.063, 1.029 and 1.328 (Table 1): every model engages more cells than the gold standard contains, and the excess clears the noise floor of OAI (.040) for every model except `deepseek-v4.1-flash` (1.029; rerun 1.046).

The direction is shared across models; the exit is not. `qwen3.8-max-0902` and `ZHIPU/GLM-5.3-FlashX` send the surplus to NEUTRAL (116 and 103 NEUTRAL predictions against 12 gold NEUTRAL cells), whereas `kimi-k3` and `deepseek-v4.1-flash` send it to ACCEPT or REJECT (51 and 30 gold-NA cells), and `kimi-k3` hardly uses NEUTRAL at all (7 predictions). These are single runs of discrete labels: switching thinking

Table 1: Bare-prompt baselines across models on IPBench (all 516 cells), pooled over the three propositions. Four-way: accuracy, macro-F1 and Cohen's κ against the single-annotator gold; activation: precision, recall, OAI and FAS; stance: accuracy and κ on gold-activated cells. Each row is a single run (not replicated); per-model tables with per-proposition breakdowns are in the appendix. n is the number of cells with a valid label.

				Four-way			**Activation**				**Stance**	
Model	Context	Think	n	Acc.	Macro-F1	κ	Prec.	Rec.	OAI	FAS	Acc.	κ
`qwen3.8-max-0902`	bare	off	516	.702	.559	.479	.673	.828	1.230	.455	.517	.326
`qwen3.8-max-0902`	bare	on	516	.529	.507	.361	.487	.994	2.040	.749	.649	.439
`kimi-k3`	bare	off	516	.783	.548	.575	.714	.759	1.063	.473	.661	.442
`kimi-k3`	bare	on	516	.636	.556	.442	.561	.931	1.661	.676	.649	.443
`deepseek-v4.1-flash`	bare	off	516	.754	.572	.525	.721	.741	1.029	.394	.557	.354
`deepseek-v4.1-flash`	bare	on	516	.570	.526	.396	.509	.994	1.954	.752	.684	.478
`ZHIPU/GLM-5.3-FlashX`	bare	on	516	.715	.577	.512	.649	.862	1.328	.551	.621	.411
`glm-5.3-prime`	bare	on	508*	.679	.590	.498	.589	.953	1.618	.693	.706	.512

Table 2: Where the four-way errors go, per model (main setting, single run; thinking on for the two GLM models, off otherwise). Each column counts cells by gold label → predicted label; the largest off-diagonal cell of each confusion matrix is in bold. Gold has 12 NEUTRAL cells.

Model	n	Errors	NA → NEUTRAL	NA → ACCEPT/REJECT	ACCEPT/REJECT → NEUTRAL	ACCEPT/REJECT → NA	ACCEPT ↔ REJECT	pred. NEUTRAL
`qwen3.8-max-0902`	516	154	**59**	11	51	24	3	116
`kimi-k3`	516	112	2	**51**	5	33	9	7
`deepseek-v4.1-flash`	516	127	20	**30**	24	40	7	50
`ZHIPU/GLM-5.3-FlashX`	516	147	**63**	18	36	19	3	103
`glm-5.3-prime`	508	163	**94**	19	37	7	4	141

on moves every model's operating point, recall up and precision down (Appendix M), so no table ranks the models.

5 Analysis

Prior Knowledge. Section 4 shows that every model judges passages unrelated to P (gold NA) as engaging it. Passage 001 (孟子見梁惠王) speaks of humaneness, righteousness and profit and never of law, its gold label for P_2 is NA, and `qwen3.8-max-0902` answers NEUTRAL while noting in its rationale that the text nowhere mentions "law". Where does the added "law" come from? We test two forms of prior knowledge, the two sides (premise and hypothesis) of the attestation bias described by McKenna et al. (2023), in which a model that remembers a statement as true judges that a premise entails it whatever the premise says: on the text side, the model's memory of the author, which attributes Mencius's doctrine to a passage that does not express it (Skinner, 1969); on the proposition side, the orthodoxy of the proposition, P_1 and P_2 being the dominant positions of the Confucian tradition. The structure of the comparison itself, a third candidate, is not tested here.

Design. The model's prior is measured first, *without the text*, and then set against its main-setting judgements of Section 4; only the activation layer is scored, since neither intervention changes the gold at that layer. Readings and thresholds were committed before the runs (Appendix L; prompts in Appendix N): decisions use gold NA cells only, a difference must exceed the noise floor, an effect counts only if at least three of the five models (the four reference models and `glm-5.3-prime`) show it, and any group under 20 cells is reported as counts only.

Text side: memory of the author does not raise false engagement. Each model names the source of each passage and says, in two wordings, whether that source holds P; the false-engagement rate is then compared between the gold-NA cells whose source the model "remembers a stance" for and those it assigns "no stance", stratified by proposition (Mantel–Haenszel risk ratio with its interval above 1, and a rate difference above the noise floor of .053). Table 21 (Appendix O): no model shows the effect. Every computable risk-

ratio interval contains 1, and on the passages whose source the model cannot name at all the false-engagement rate (.20–.36) is no lower than its overall rate.

Proposition side: orthodoxy has an unstable, model-dependent effect. P_1 and P_2 are replaced by the other side of the same comparison, P_1' and P_2' (Appendix F); the gold at the activation layer is unchanged, so the same cells are rerun with the same prompt, and all five models rank the original proposition above its contrary when asked without any text (the pre-registered manipulation check). A model shows the effect when more gold-NA cells are engaged only under the original than only under the contrary proposition (one-sided exact McNemar test, $p < .05$) and its false-engagement rate drops by more than the noise floor of .015 (readings in Table 6, Appendix L). Table 22 (Appendix O) gives the activation-layer metrics under both propositions: two models show the effect, short of the three required. `deepseek-v4.1-flash` and `ZHIPU/GLM-5.3-FlashX` lower their OAI ($1.07 \to .93$ and $1.47 \to 1.38$), `qwen3.8-max-0902` does not move, `kimi-k3` moves the other way ($1.19 \to 1.34$), and `glm-5.3-prime` moves without meeting the criterion.

Summary. Neither form of prior knowledge accounts for the volume of over-engagement: removing the author leaves it unchanged, and reversing the proposition relabels it rather than withdrawing it, the false engagements that were ACCEPT or NEUTRAL under the original proposition becoming REJECT under the contrary one (post-hoc, single runs; Appendix L). The models read correctly which side of a comparison a passage stands on; what they do not check is whether the passage makes the comparison at all, which is the boundary that the position list Π_P of Section 2.2 draws. Both experiments are single runs with small groups, so the result is that no evidence was found, not that an effect is excluded.

6 Conclusion

We have defined the ideational proposition, a hypothesis that is exogenous and abstract and ideational, and the judgement TWPI that relates a text to it in two steps, engagement and stance. IPBench instantiates the judgement on 516 cells labelled by two experts under released conventions, with a protocol that scores the two steps separately and measures over-engagement as a quantity of its own. Under one frozen bare prompt, four models put most of their errors on the engagement step, every one of them engaging more cells than the gold standard contains, with the surplus exiting through model-specific labels. Two pre-registered experiments find little and no consistent effect of prior knowledge, whether memory of the author or orthodoxy of the proposition, on that surplus: the models read which side of a comparison a passage stands on but do not check whether it makes the comparison at all. Applied over a fixed set of propositions, the judgement maps every text to a point of a shared proposition space; IPBench is not a new judge and its tables are not a ranking, but a testbed for the operation that such a space presupposes.

Two limitations bound these claims. The gold standard is the first of two expert annotations that agree at $\kappa = .92$ (Section 3.2); on the cells where the experts differ, and where a convention is contestable, a model that reads the cell otherwise is scored as wrong, and every κ reported for a model is model–gold agreement. The evaluation is small in every dimension: three propositions on one corpus selected for relevance to them, four models under one template, single runs of discrete labels with one rerun, so that no table ranks the models, the noise floor is coarse, and generalisation to other propositions, corpora or models is untested.

AI Use Statement

Models as the object of study. The systems evaluated in Sections 4 and 5 are generative models: `qwen3.8-max-0902`, `kimi-k3`, `deepseek-v4.1-flash`, `ZHIPU/GLM-5.3-FlashX` and `glm-5.3-prime`, called through one OpenAI-compatible interface on 24–25 September 2026 with the settings of Section 4; their outputs are data, not assistance. The position lists Π_P of Appendix J were generated by a locally deployed Qwen3.8-27B-FP8 from the proposition wording (Section 2.2) and are released verbatim.

Disclosure. In this work, we used generative AI tools (Claude Code, Anthropic) to aid and polish the writing (the Chinese drafts of Sections 3–5 and the quoted passages were translated into English by the tool, and the text was edited for readability; translations of quoted passages are marked as machine-assisted where they appear), for retrieval and discovery (identifying and summarising related work and formatting references; every bibliographic entry added with the tool's help was checked against the publishers' pages before inclusion), and to draft sections of the paper (Sections 3–5 and the appendices were drafted by the tool from the authors' notes, the plan files and the result files, and revised by the authors). We have not used generative AI tools for research ideation or execution: the task definition, the propositions, the annotation protocol, the hypotheses, decision rules and thresholds of the pre-registration plans, and the design of the experiments are the authors'; the tool wrote the plan files, the calling scripts, the evaluation scripts and the table generators from the authors' specifications, and the authors reviewed and froze them before the runs they govern. We have not used generative AI tools to generate synthetic data sets, to formulate mathematical claims or to assist in proving them, or to clean or reformat the data set. Additionally, we used generative AI tools for creating or editing software code, formatting references, creating the schematic figure, and suggesting the structure of the paper. We have reviewed all AI-assisted work: the authors reviewed and annotated every section and directed each revision, the tables are generated directly from the result files by released scripts, and every number in the text is traced in the project record to its source file. The gold labels were produced by the two experts without generative AI and before the model runs reported here, from the protocol's task statement, label set and conventions (Appendix G). We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics Statement

IPBench consists of passages from a published scholarly compilation of classical Chinese philosophical writing; the texts are historical and involve no living subjects and no personal data. Annotation was performed by two researchers with training in intellectual history, as part of their research work (Section 3.2).

Two risks deserve naming. First, the ideational propositions and the semantic conventions used to operationalise them are interpretive choices made by an annotator working in one scholarly tradition; a different tradition could reasonably draw the engagement boundary elsewhere. We release the conventions explicitly rather than leaving them implicit in the labels, so that this is contestable rather than hidden. Second, a system that automatically attributes ideological commitments to texts and authors can be misused to produce authoritative-sounding characterisations of what a tradition "believed". Our own results argue against such use at present: the dominant failure mode is asserting engagement where there is none.

Reproducibility Statement

The corpus, the three propositions with their semantic conventions, all 516 gold labels of the single annotator, the evaluation template of Appendix J with the position lists Π_P, the prompts of Appendix N, the pre-registration documents and deviation logs of Appendix L, and the per-cell predictions of every setting and every rerun reported here will be released with the camera-ready version; the gold labels are the first expert's annotation (Section 3.2).

All reported runs call the models through one OpenAI-compatible interface at temperature 0 and seed 0, one stateless call per cell with no system prompt, under parsing and invalid-cell rules fixed before the run (Section 4). The hypotheses of the baseline and of the two controlled experiments, with their decision rules and thresholds, were written down and frozen before the full runs they govern, and every deviation was logged (Section 5). Two caveats are part of the record: identical settings do not make the service deterministic, so each main setting was run twice independently and the per-metric noise floors of Table 20 are taken from those reruns; and the position lists Π_P were generated by a locally deployed Qwen3.8-27B-FP8 from the proposition wording (Section 2.2) and are released verbatim rather than regenerated.

References

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. From arguments to key points: Towards automatic argument summarization. In *Proceedings of ACL*, pp. 4029–4039, 2020a.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. Quantitative argument summarization and beyond: Cross-domain key point analysis. In *Proceedings of EMNLP*, pp. 39–49, 2020b.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. Abductive commonsense reasoning. In *International Conference on Learning Representations (ICLR)*, 2020.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*, pp. 632–642, 2015.

Robert B. Brandom. *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Harvard University Press, 1994.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges: Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment (MLCW 2005)*, volume 3944 of *Lecture Notes in Computer Science*, pp. 177–190. Springer, 2006.

Roni Friedman, Lena Dankin, Yufang Hou, Ranit Aharonov, Yoav Katz, and Noam Slonim. Overview of the 2021 key point analysis shared task. In *Proceedings of the 8th Workshop on Argument Mining*, pp. 154–164, 2021.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4878–4887, 2017.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30): e2305016120, 2023. URL `https://www.pnas.org/doi/10.1073/pnas.2305016120`.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of NAACL-HLT (Short Papers)*, pp. 107–112, 2018.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. A survey on stance detection for mis- and disinformation identification. *Findings of NAACL*, 2022.

Fritz Heider. *The Psychology of Interpersonal Relations*. Wiley, 1958.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021.

Ulf Hlobil and Robert B. Brandom. *Reasons for Logic, Logic for Reasons: Pragmatics, Semantics, and Conceptual Roles*. Routledge, 2024.

Jerry R. Hobbs, Mark E. Stickel, Douglas E. Appelt, and Paul Martin. Interpretation as abduction. *Artificial Intelligence*, 63(1–2):69–142, 1993.

Guantao Jin and Qingfeng Liu. 觀念史研究：中國現代重要政治術語的形成 *[Studies in the History of Ideas: The Formation of Key Modern Chinese Political Terms]*. 法律出版社 [Law Press China], Beijing, 2009.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025.

Walter Kintsch. The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95(2):163–182, 1988.

Walter Kintsch. *Comprehension: A Paradigm for Cognition*. Cambridge University Press, 1998.

Walter Kintsch and Teun A. van Dijk. Toward a model of text comprehension and production. *Psychological Review*, 85(5):363–394, 1978.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. AbstentionBench: Reasoning LLMs fail on unanswerable questions. In *Advances in Neural Information Processing Systems*, 2025.

Reinhart Koselleck. *Futures Past: On the Semantics of Historical Time*. Columbia University Press, 2004. Translated by Keith Tribe.

Michal Kosinski. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121, 2024.

Matthew D. Lieberman, Johanna M. Jarcho, and Junko Obayashi. Attributional inference across cultures: Similar automatic attributions and different controlled corrections. *Personality and Social Psychology Bulletin*, 31(7):889–901, 2005.

Zipeng Ling, Shuliang Liu, Yuehao Tang, Junqi Yang, Shenghong Fu, Seonil Son, Chen Huang, Kejia Huang, Yao Wan, Zhichao Hou, and Xuming Hu. LLM abstention can be a prompt artifact, in addition to genuine uncertainty. *arXiv preprint arXiv:2507.16199*, 2025.

Marco Lippi and Paolo Torroni. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology*, 16(2):1–25, 2016.

Chao-Lin Liu, Guantao Jin, Qingfeng Liu, Wei-Yun Chiu, and Yih-Soong Yu. Some chances and challenges in applying language technologies to historical studies in Chinese. *International Journal of Computational Linguistics and Chinese Language Processing*, 16(1–2): 27–46, 2012.

Arthur O. Lovejoy. *The Great Chain of Being: A Study of the History of an Idea*. Harvard University Press, 1936.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of ACL*, pp. 3428–3448, 2019.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. Sources of hallucination by large language models on inference tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2758–2774, 2023.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of SemEval*, pp. 31–41, 2016.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of *SEM*, pp. 180–191, 2018.

Xin Quan, Jiafeng Xiong, Marco Valentino, and André Freitas. Inferring latent intentions: Attributional natural language inference in LLM agents. *arXiv preprint arXiv:2601.08742*, 2026.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of ACL*, pp. 784–789, 2018.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. Neural theory-of-mind? on the limits of social intelligence in large LMs. In *Proceedings of EMNLP*, pp. 3762–3780, 2022.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*, 2024.

Quentin Skinner. Meaning and understanding in the history of ideas. *History and Theory*, 8(1):3–53, 1969.

Christian Stab and Iryna Gurevych. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of EMNLP*, pp. 46–56, 2014.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: A large-scale dataset for fact extraction and VERification. In *Proceedings of NAACL-HLT*, pp. 809–819, 2018.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html`.

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

Teun A. van Dijk. *Macrostructures: An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition*. Lawrence Erlbaum, Hillsdale, NJ, 1980.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of ACL*, pp. 9440–9450, 2024.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL-HLT*, pp. 1112–1122, 2018.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4913–4922, 2021.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.