

Ethical Imprints in LLMs: From Corpus Ethics to ICL, SFT, and Their Interaction on Deception Dilemmas

Jiafeng Wang¹ Jiale Chen² Zheng Luo¹ Xinyi Yan³ Lingpeng Kong^{1*}

¹The University of Hong Kong ²Hohai University

³Nanjing University of Science and Technology

{wangjf2024, dingl120}@connect.hku.hk lpk@cs.hku.hk

jialechen@hhu.edu.cn yanxinyi@njust.edu.cn

Abstract

Large language models make consistent moral judgments, but whether these tendencies correspond to the moral makeup of public web text in the same language is not well understood. We test this correspondence for deception dilemmas across Chinese and English in a four-layer empirical study. We use DeepSeek-R1 for ICL probing, Llama-3.1-8B and Qwen2.5-7B for fine-tuning, and GPT-4o for cross-model replication. In both languages, roughly 60% of web text favors utilitarian reasoning (59.8% vs. 62.3%), inconsistent with a purely sociolinguistic explanation. In-context learning shifts model decisions sharply in low-resistance scenarios (Hostage Crisis, $V = 0.411$, $p = 0.001$). In a high-resistance scenario (Organ Allocation), one DeepSeek-R1 deployment showed near-zero ICL effects ($V = 0.067$) at small sample size, but a larger replication on another deployment showed a large effect ($V = 0.55$ to 0.69), indicating deployment-specific rather than stable resistance. Three findings stand out for multilingual alignment. First, fine-tuning on deontological data reverses direction in English (Llama-EN Deon $P(A)$ rises from $\sim 0\%$ to 88%), showing that SFT value transfer is not language-neutral. Second, corpus proportions do not carry over to forced choices. The English Llama-3.1-8B base model gives $P(A)$ near zero even though 62.3% of the web text is utilitarian. Third, apparent moral damping in high-resistance scenarios can reflect sample size and deployment variability rather than a stable property of the model or scenario. These results show that corpus moral distributions correspond to model behavior in many conditions, but language, post-training, model architecture, and deployment variability each introduce points where this correspondence breaks.

1 Introduction

Large language models show consistent moral preferences across repeated runs, yet these preferences

shift under different prompts. Prior work has mostly addressed this through architecture design, alignment techniques (RLHF, DPO), or prompting strategies such as few-shot learning (Brown et al., 2020) and chain-of-thought (Askell et al., 2021; Wei et al., 2023). Less work has asked what these preferences track. We call this stable tendency the model’s ethical imprint. It is a measurable, repeatable tendency for a model to favor one moral framework, either utilitarianism or deontology, on deception dilemmas under a fixed protocol. This paper studies what shapes and shifts this imprint: corpus balance, language, post-training, ICL, and SFT.

We ask whether a model’s moral tendencies track its sociolinguistic context or the statistical makeup of web text in that language, and where this correspondence holds and where it breaks. We test whether the distribution of ethical frameworks in the web text of a given language corresponds to model moral preferences, beyond what language or culture alone would predict. Our experiments test for consistent distributional alignment across languages, models, and intervention types. We do not claim that specific corpora caused specific model behaviors. Three patterns stand out. Utilitarian shares in our sampled Chinese and English corpora are both near 60%, so language alone is unlikely to explain the corpus gap. Equal-sized English deontological SFT data does not suppress utilitarian tendencies. It reverses them instead, suggesting that deontological SFT outcomes depend on the model’s pretraining language. Neutral chain-of-thought suppresses utilitarian responses in Chinese: Neutral scores below Control in all six cells. In English it promotes them instead: Neutral is at or above Control in five of six cells. This asymmetry lines up with the language of each setting, not with the direction of the examples used.

To test this systematically, we focus on deception, where utilitarianism and deontology disagree

* Corresponding author.

most sharply. We build a four-layer verification framework across Chinese and English. Each layer tests a different aspect. The corpus layer measures utilitarian and deontological proportions in large-scale web corpora (CCI 4.0 and FineWeb) as representative samples of Chinese and English web text, not as replicas of any model’s training set. The mechanism layer uses DeepSeek-R1 with ICL to test whether models reflect the moral framework in their input. The model layer uses Llama-3.1-8B and Qwen2.5-7B with SFT to test whether these distributions can be encoded into parameters. The interaction layer examines how parametric encoding and context interact on the fine-tuned models.

Contributions (1) We conduct a systematic four-layer empirical investigation (corpus, mechanism, model, interaction) replicated across Chinese/English, linking language-level web-text moral distributions to model preferences.

(2) Cross-lingual corpus annotation shows utilitarian proportions near 60% in both languages, inconsistent with a language-only account.

(3) Deontological SFT produces a language-dependent asymmetry: it drives $P(A)$ to near zero in Chinese but *reverses direction* in English (Llama-EN Deon rises from $\sim 0\%$ to 88%). Value transfer through SFT is not language-neutral.

(4) Corpus proportions do not carry over to forced choices. The English Llama-3.1-8B base model gives $P(A)$ near zero even though 62.3% of the English web corpus is utilitarian. This model is a pretrained checkpoint with no post-training, so the gap opens before any alignment step.

(5) Moral damping in Organ Allocation does not hold up. One DeepSeek-R1 deployment showed near-zero ICL effects at small sample size ($V=0.067$). A larger replication on another deployment found a much larger effect ($V=0.55$ to 0.69), and GPT-4o shows a strong effect in the same scenario ($V=0.648$). This damping depends on deployment and model, not on the scenario itself.

2 Related Work

2.1 Measuring Moral Preferences in LLMs

Human moral preferences have long been studied through frameworks such as Moral Foundations Theory (Haidt and Graham, 2007; Graham et al., 2011). Recent work probes whether LLMs show similar preferences. Several benchmarks, including the ETHICS dataset (Hendrycks et al., 2021), Delphi (Jiang et al., 2021), the Moral Machine

(Awad et al., 2018), and the Greatest Good Benchmark (Marraffini et al., 2024), find that models lean utilitarian in low-ambiguity cases (Scherrer et al., 2023). The strength of this lean varies across model families (Zheng et al., 2023) and is sensitive to context (Barabadi et al., 2026). Cross-lingual variation adds another dimension: Liu et al. (2024) reported that English- and Chinese-dominant models diverge along individualist–collectivist lines. Schramowski et al. (2022) recovered a moral direction from pretrained embeddings without any explicit moral training, suggesting these tendencies come from the data itself. However, none of these studies test whether preferences correspond to corpus-level distributions or how pretraining bias interacts with ICL or SFT, which we address here.

2.2 Cross-Lingual Moral Reasoning

Jin et al. (2025) found significant cross-lingual differences in LLM moral judgments across more than 100 languages. Our work uses a controlled design to investigate what these differences track.

2.3 ICL and SFT Effects on Values

Both ICL and SFT can redirect model behavior, but how the two interact is not well studied. On the ICL side, Wei et al. (2023) showed that large models rely more on pretrained knowledge than on example labels. Lin et al. (2024) showed that even a few well-formatted examples can activate latent aligned behaviors. Yet Chochlakis et al. (2024) found that chain-of-thought prompting (Wei et al., 2022) fails to shift subjective value judgments, pointing to hard limits on context-driven steering. On the SFT side, small high-quality datasets can change outputs (Solaiman and Dennison, 2021; Zhou et al., 2023), but this also creates risks: Qi et al. (2024) showed that just 10 adversarial examples suffice to break RLHF safety alignment. RLHF (Christiano et al., 2017; Ouyang et al., 2022) and Constitutional AI (Bai et al., 2022) set the broader training-time context. Yuan et al. (2026) found that post-RLHF reasoning chains improve honesty. This raises the question of whether reasoning-only prompts interact differently with moral preferences across languages, which we test in §5.2.

2.4 Pretraining Data Distribution and Moral Preferences

Feng et al. (2023) showed that political bias in pretraining data corresponds to measurable downstream effects in hate-speech detection and misin-

formation identification. This provides a methodological precedent for our multi-layer framework. Both Feng et al. (2023) and Schramowski et al. (2022) focused on political or safety dimensions and did not test how pretraining bias interacts with ICL or SFT. We extend this approach to moral-framework distributions with a four-layer design.

3 Methods

3.1 Scenario Selection: Why Deception?

Deception is a clear test case for comparing utilitarianism and deontology. Deontological theory treats lying as unconditionally wrong (Kant, 1785), while utilitarianism permits or requires it when the outcome justifies it (Bentham, 1789; Mill, 1863). Deception also appears frequently in varied web text. We standardize it into three scenarios with different levels of institutional resistance: Friend Referral (low), Hostage Crisis (medium-low), and Organ Allocation (high). These are inspired by the trolley problem tradition (Foot, 1967; Thomson, 1985) and represent low, medium, and high institutional constraint.

3.2 Corpus Collection and Annotation

Chinese text was drawn from CCI 4.0 (Common Crawl Chinese).¹ English text was drawn from FineWeb (Penedo et al., 2024),² specifically the base version (unfiltered general web text, not FineWeb-Edu), to avoid overestimating deontological proportions and to maintain comparability with CCI 4.0. We chose these two corpora because they are among the largest publicly available web-crawl datasets in their respective languages. CCI 4.0 overlaps substantially with the Chinese web text used to train Chinese-centric models; FineWeb is derived from Common Crawl, which forms the backbone of most English pretraining pipelines (Brown et al., 2020; Touvron et al., 2023). Neither corpus is the verified training set of any model tested here; we return to this limitation in §8.

After keyword filtering, 2,000 candidates per language were sampled. Annotation used DeepSeek-R1 (Volcengine ARK API) with semantically equivalent bilingual prompts verified by back-translation. Quality control included three rounds of independent annotation, confidence-threshold filter-

ing ($\text{Relevance_Decision_Certainty} \geq 0.8$, $\text{Framework_Mapping_Certainty} \geq 0.7$), Fleiss' kappa computation, and removal of samples where all three rounds disagreed. Final counts: 321 Chinese (192 utilitarian, 129 deontological) and 199 English (124 utilitarian, 75 deontological). Fleiss' kappa: $\text{is_relevant } \kappa=0.76$ (substantial); $\text{moral_framework } \kappa=0.52$ (moderate). The lower framework kappa reflects real ambiguity at the boundary between frameworks (Artstein and Poesio, 2008).

DeepSeek-R1 plays two roles: corpus annotator and ICL test model. Three points argue against circular reasoning. First, R1 matched human annotation most closely among three candidate models. Second, model-layer SFT used Llama-3.1-8B and Qwen2.5-7B, which are architecturally independent of R1, and both showed theoretically expected shifts on R1-labeled data. Third, the mechanism layer measures directional changes in $P(A)$ between Deon and Util contexts. Directional effects are much less sensitive to systematic annotator bias than absolute label accuracy.

3.3 Four-Layer Experimental Design

Corpus layer. We tabulate annotated utilitarian and deontological proportions and run a chi-square test comparing Chinese and English.

Mechanism layer. Using DeepSeek-R1, we run four-context ICL experiments (Deon, Util, Neutral, Control) on three standardized scenarios, 30 samples each, at $T=0.2$ and $T=0.8$, with $P(A)$ as the metric. Deon, Util, and Neutral contexts each provide few-shot chain-of-thought examples that differ only in moral direction. Neutral provides reasoning without a moral stance, to separate the reasoning process from moral content. Control has no examples. GPT-4o replicates the same protocol independently (§5.3).

Model layer. We build SFT training sets (Chinese: 192 utilitarian, 129 deontological; English: 124 utilitarian, 75 deontological) and use LoRA (Hu et al., 2022) to fine-tune Llama-3.1-8B (Meta AI, 2024) and Qwen2.5-7B ($r=16$, $\alpha=16$, $\text{lr}=2 \times 10^{-4}$, 5 epochs). Conditions: Base, Util_ZH, Deon_ZH, Util_EN, Deon_EN.

Interaction layer. We run four-context ICL experiments on the SFT-tuned models and analyze how parametric encoding and context interact.

¹CCI 4.0: <https://huggingface.co/datasets/BAAI/CCI4-Data>

²FineWeb: <https://huggingface.co/datasets/HuggingFaceFW/fineweb>

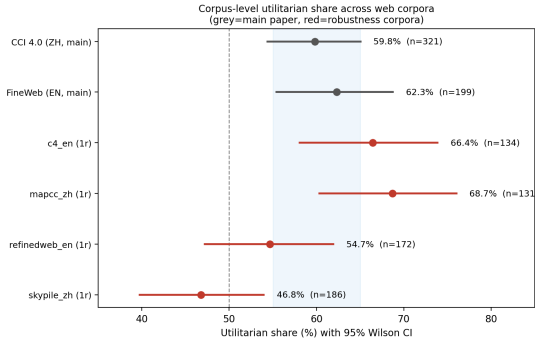


Figure 1(Robustness): Forest plot of utilitarian share across six corpora. Points show the estimated share; horizontal lines show 95% Wilson confidence intervals. Gray marks the main-experiment corpora (CCI 4.0, FineWeb); red marks the four robustness corpora; the dashed vertical line marks the 50% reference.

4 Corpus Layer: Distributional Baseline

Chinese and English moral framework distributions are nearly identical (Table 1). A chi-square test shows no significant difference. The 2.5 pp utilitarian gap is too small to support a language-only story or a single-dataset artifact. We use this near-identical distribution as a matched control, not as causal evidence. Because the utilitarian share is almost the same in both corpora, a large divergence in model behavior would not be explained by the corpus’s ethical makeup. This isolates the roles of language and post-training. We do not claim that the corpus distribution explains model behavior. The 95% confidence intervals for the two utilitarian shares overlap substantially (Chinese 59.8%, [54.4%, 65.0%], $n=321$; English 62.3%, [55.4%, 68.8%], $n=199$). This further supports the distributional match.

To test whether this distribution is an artifact of the two specific datasets, CCI 4.0 and FineWeb, we ran the same annotation protocol on four additional large-scale web corpora. These include the public training sets of three open-source LLMs (C4/T5, RefinedWeb/Falcon, SkyPile/Skywork) and one independent Chinese web corpus (MAP-CC). Utilitarian shares across all six corpora range from 46.8% to 68.7%. The utilitarian lean replicates in most corpora (four of six have confidence intervals excluding 50%), and the remaining variation falls between corpora rather than between languages. This shows that the near-identical distribution is not an inevitable result of a shared Common Crawl origin, but a testable empirical pattern (Figure 1(Robustness); full per-corpus counts in Appendix F).

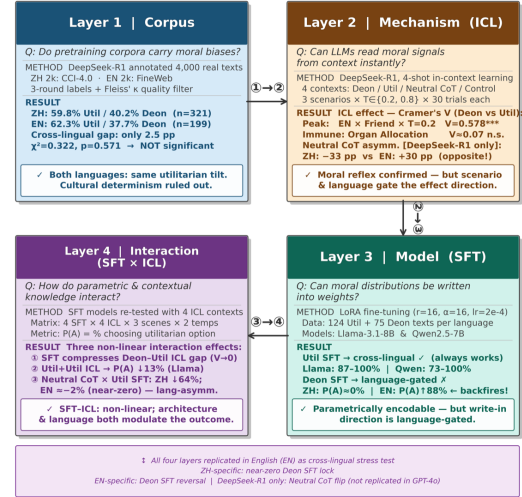


Figure 1: Four-layer verification framework. Each layer tests a different aspect of the correspondence between corpus-level ethical distributions and model moral behavior: Corpus Layer (distributional baseline), Mechanism Layer (ICL context sensitivity), Model Layer (SFT parametric encoding), Interaction Layer (SFT $\times$ ICL dynamics).

Figure 1 outlines the four-layer framework used in the mechanism, model, and interaction sections below.

5 Mechanism Layer: Context Sensitivity

5.1 Scene-Dependent Moral Damping

ICL effects on moral decisions depend strongly on the scenario, and this pattern holds across both languages. Hostage Crisis shows the strongest context modulation. In Chinese, Organ Allocation shows near-zero response to any intervention across all temperatures on DeepSeek-R1 ($V=0.067, p=0.604$ at $T=0.2$; $V=0.070, p=0.588$ at $T=0.8$), a pattern we call moral damping (near-zero ICL effect size). This damping, however, is not robust. The original Volcengine Ark endpoint (deepseek-r1-250528) is no longer available, so we reran this check at $n=100$ on the openly hosted DeepSeek-R1-0528, the same released weights on a different deployment. On this deployment, Chinese Organ Alloca-

	Total	Valid	Util	Deon
ZH (CCI)	2000	321	192 (59.8%)	129 (40.2%)
EN (FineWeb)	2000	199	124 (62.3%)	75 (37.7%)
$\chi^2=0.322, p=0.571$			$V=0.024$ (n.s.)	

Table 1: Annotation statistics. Utilitarian gap 2.5 pp; $\chi^2=0.322, p=0.571, V=0.024$ (n.s.).

Scenario	T	ZH D	ZH U	ZH V	EN D	EN U	EN V
Friend	0.2	26.7	56.7	.270*	30.0	90.0	.578★
Friend	0.8	23.3	50.0	.242	60.0	80.0	.182
Hostage	0.2	30.0	60.0	.268*	53.3	80.0	.247
Hostage	0.8	16.7	60.0	.411**	50.0	80.0	.280*
Organ	0.2	40.0	50.0	.067	56.7	90.0	.339**
Organ	0.8	30.0	40.0	.070	70.0	83.3	.118

Table 2: Cramér’s V across Chinese and English ICL (Deon vs. Util). ★ $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; unmarked = n.s. All $P(A)$ in %.

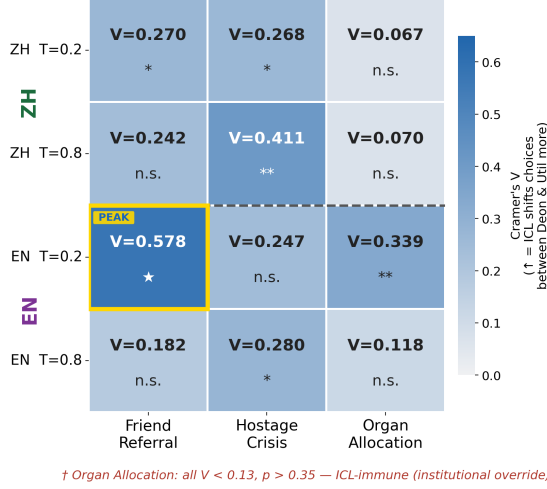


Figure 2: ICL moral reflection effect heatmap (Cramér’s V , Deon ICL vs. Util ICL). Each cell = effect size of the four-context ICL on $P(A)$; darker = stronger intervention. ★ $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; n.s. = not significant.

tion showed a large context effect (Deon vs. Util: $V=0.55$ at $T=0.2$, $V=0.69$ at $T=0.8$, both $p < 1e-14$, with $P(A)$ rising from 0.20 in the deontological context to 0.76 in the utilitarian context). Because the original deployment is no longer accessible, we cannot rerun it at $n=100$, so we cannot tell whether this reflects small-sample noise or a genuine deployment difference. Either way, it shows that the damping depends on deployment and model, rather than being a stable immunity built into the scenario. In English, however, Organ Allocation shows a significant ICL effect only at $T=0.2$ ($V=0.339$, $p=0.009$), while $T=0.8$ does not reach significance ($V=0.118$, $p=0.360$), indicating that moral damping in this scenario is partially language-specific on DeepSeek-R1. Under GPT-4o, damping disappears entirely in both languages (§5.3).

Table 2 reports Cramér’s V for every scenario, language, and temperature; Figure 2 visualises the same effect sizes as a heatmap.

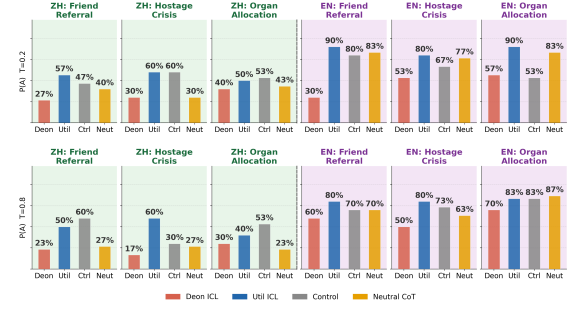


Figure 3: Full Chinese/English ICL context intervention results (DeepSeek-R1, 4 contexts $\times$ 3 scenarios $\times$ 2 temperatures). Left three columns: Chinese (ZH); right three: English (EN). Top row: $T=0.2$; bottom: $T=0.8$.

5.2 Language Boundary Condition of Neutral CoT

Neutral chain-of-thought reasoning has opposite effects in DeepSeek-R1 depending on the language. In Chinese, neutral reasoning lowers $P(A)$ below Control in all conditions, suggesting the model’s analytical process draws on deontological tendencies. In English, the direction reverses: $P(A)$ rises above Control in most conditions. What matters is not whether reasoning occurs, but what it draws on. The neutral context acts as a reasoning-only control. Its reversal across languages points to different moral tendencies in analytical web text. GPT-4o did not show this pattern, so the effect may be specific to DeepSeek-R1’s training.

Figure 3 reports full results across 24 cells (four contexts $\times$ three scenarios $\times$ two temperatures).

5.3 Cross-Model Validation: GPT-4o Replication

We ran the same four-context ICL experiment with GPT-4o (OpenAI, 2024) (checkpoint: gpt-4o-2024-11-20) on all three scenarios in both languages ($n=30$ per cell, $T=0.2$ and $T=0.8$). ICL polarity effects replicate fully: all six scenario groups are significant ($V=0.431$ to 0.978 , all $p < 0.0001$). Context sensitivity is not specific to one model or provider. GPT-4o shows no damping in either language on Organ Allocation (Chinese: $V=0.648$; English: $V=0.788$), unlike the near-zero effect on DeepSeek-R1’s original deployment. This fits the deployment-specific pattern noted above, not a general immunity in the scenario itself. Baseline tendencies also differ: GPT-4o Chinese Friend Referral Control $P(A)$ is 0%, compared to around 50% for DeepSeek-R1.

Figure 3(Supplement) summarizes the GPT-4o

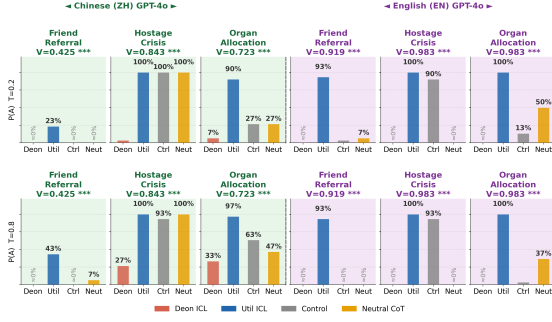


Figure 3(Supplement): GPT-4o ICL moral context experiments (Chinese/English, $n=30/\text{cell}$, 4 contexts $\times$ 3 scenarios $\times$ 2 temperatures). All scenarios significant ($V=0.431$ to 0.978); Organ Allocation damping absent in both languages, confirming that damping is specific to Chinese DeepSeek-R1, not a cross-model or cross-lingual universal.

replication panel.

6 Model Layer: Parametric Encoding

6.1 Utilitarian SFT: Cross-Lingual and Cross-Architecture

Utilitarian SFT produces significant $P(A)$ gains in both languages and both architectures (Table 3, Figure 4). Differences in gain size are largely due to different starting points in the base models; both converge after fine-tuning. One exception: ZH-Qwen Organ Allocation at $T=0.2$ shows Util SFT $P(A)$ slightly below Base, echoing the near-zero ICL effect seen in this scenario.

6.2 Deontological SFT: Language-Specific Asymmetry

Deontological SFT diverges sharply by language. Chinese runs reach near-zero $P(A)$; English runs reverse: Deon SFT raises $P(A)$ instead of lowering it (Table 3). This reversal holds across temperatures.

An equal-sample control (zh_deon_75: 75 Chinese deontological samples) also achieves $P(A)$

Model	Base	Util SFT	Deon SFT
Llama-ZH	$\approx 1.1\%$	$88.9\% \uparrow$	$\approx 0\%$
Qwen-ZH	$\approx 60\%$	$\approx 79\% \uparrow$	$\approx 2\% \downarrow$
Llama-EN	$\approx 0\%$	$\approx 92\% \uparrow$	$88\% \uparrow \star$
Qwen-EN	$\approx 78\%$	$\approx 99\% \uparrow$	$84\% \uparrow \star$
zh_deon_75	$\approx 1.1\%$	—	$\leq 10\%$

Table 3: SFT results ($P(A)$, @ $T=0.2$, 3-scenario mean). English deontological SFT reversal: Llama-EN Deon rises from $\sim 0\%$ to 88% ; Qwen-EN similarly reverses. $\star$ marks direction reversal.

at or below 10%, ruling out sample size as the explanation (Table 4, Figure 5). We do not have a confirmed explanation for this reversal. Three post-hoc hypotheses fit the data but remain untested. First, the English pretrained model may not represent absolute rules well in this answer format, so training toward deontology pushes it the other way. Second, with 62.3% utilitarian pretraining data, gradient updates toward deontology may drift back toward the utilitarian majority (Zhou et al., 2023; Hu et al., 2022). Third, the SFT classification format may work differently on English and Chinese pretrained representations (Askell et al., 2021; Lin et al., 2024). Distinguishing these requires activation-level analysis. As a pilot test, we ran a preliminary causal check on whether this reversal is carried by a specific, movable set of middle-layer representations. We used activation patching on Llama-3.1-8B. We transplanted the middle-layer residual-stream activations from the Chinese Deon-SFT model into the English Deon-SFT model, at the layers identified by a layer-by-layer sweep. This single intervention pulled the English utilitarian rate $P(A)$ back toward the near-deontological behavior seen in Chinese: Organ Allocation $0.66 \rightarrow 0.18$, Hostage Crisis $0.72 \rightarrow 0.28$, Friend Referral $0.72 \rightarrow 0.38$. The effect is concentrated around layers 7 to 18. This shows that the reversal is not a surface artifact but is encoded in a specific, causally movable set of middle-layer

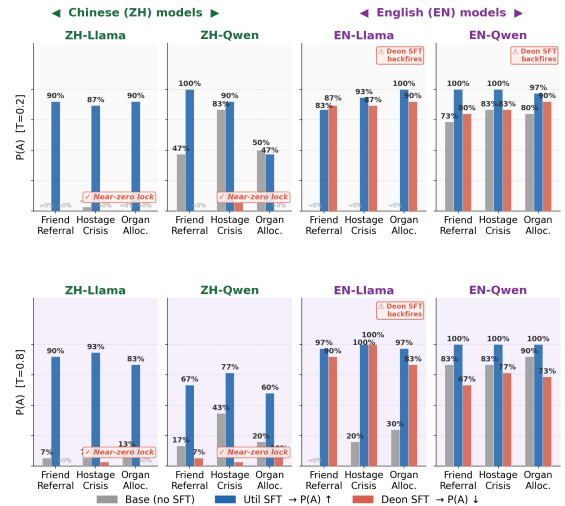


Figure 4: SFT parametric moral encoding results. $P(A)$ = Control-group utilitarian response rate (30 samples/scenario). Utilitarian SFT effective cross-linguistically and cross-architecturally; deontological SFT language-specific. Near-zero anchoring in Chinese, full directional reversal in English.

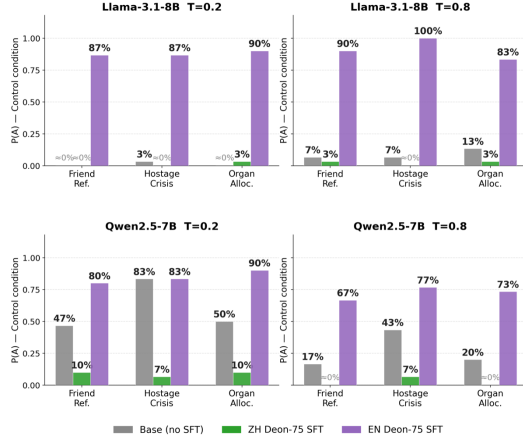


Figure 5: Equal-sample Deon SFT control (75 samples, Control-group $P(A)$). Grey = Base; green = ZH Deon-75; purple = EN Deon-75. Chinese achieves near-zero anchoring; English fails to suppress utilitarian tendencies and reverses.

representations.

7 Interaction Layer: Mechanistic Boundaries

7.1 Resolution Suppression

SFT has two separable effects. First, the gap in $P(A)$ between Deon and Util ICL contexts narrows (we call this resolution suppression). Second, $P(A)$ shifts toward the extreme matching the training framework (level anchoring). ZH-Qwen-Deon achieves both. EN-Qwen-Deon partially achieves the first but reverses the second (Table 5). Parametric encoding and behavioral locking are distinct mechanisms.

After SFT, the Deon/Util $P(A)$ gap narrows in most scenarios, but sensitivity to the four ICL contexts is preserved. Under Util SFT, the two architectures respond differently to the Neutral context. ZH-Qwen Neutral CoT shifts from suppressing util-

Scenario T	Llama			Qwen		
	ZH-129	ZH-75	EN-75	ZH-129	ZH-75	EN-75
Friend 0.2	0.0	0.0	86.7	0.0	10.0	80.0
Friend 0.8	0.0	3.3	90.0	6.7	0.0	66.7
Hostage 0.2	0.0	0.0	86.7	6.7	6.7	83.3
Hostage 0.8	3.3	0.0	100.0	3.3	6.7	76.7
Organ 0.2	0.0	3.3	90.0	0.0	10.0	90.0
Organ 0.8	0.0	3.3	83.3	10.0	0.0	73.3

Table 4: Equal-sample deontological SFT control (75 samples). Chinese Deon-75 achieves near-zero anchoring; English Deon-75 fully reverses. All values $P(A)$ in %.

Model	Scenario	Base gap	Post-SFT gap
Qwen-ZH Util	Hostage	+90pp★	+23.3pp *
Qwen-ZH Util	Organ	+86.7pp★	+13.3pp n.s.
Llama-EN Util	Hostage	+63.3pp★	−6.7pp n.s.
Qwen-EN Util	Friend	+53.3pp★	0pp (ceiling)

Table 5: Resolution suppression. Deon vs. Util $P(A)$ gap (Base vs. SFT). All SFT gaps narrowed ($*p<0.05$); full data in Appendix D.

itarian tendencies to amplifying them, suggesting deep internalization. ZH-Llama’s high $P(A)$ drops under neutral reasoning, suggesting shallow internalization. Separating language mismatch from other factors requires further work.

Table 5 quantifies resolution suppression (Deon vs. Util $P(A)$ gap before and after SFT).

7.2 Defensive Contraction

When a Util SFT model receives Util ICL examples, $P(A)$ sometimes falls below the no-context Control baseline. We refer to this pattern as defensive contraction. The effect is strongest in ZH-Llama. At $T=0.8$, the Util SFT plus Util ICL condition falls below Control by 63.3 percentage points in Friend Referral, 40.0 pp in Hostage Crisis, and 50.0 pp in Organ Allocation. At $T=0.2$, Friend Referral and Hostage Crisis show contractions of 10.0 to 26.7 pp. ZH-Qwen shows the effect in Friend Referral ($T=0.2$: −23.3 pp; $T=0.8$: −13.3 pp). EN-Llama shows a mild version (−6.7 to 13.3 pp). EN-Qwen shows none.

This pattern is consistent with several recent findings. Redundant few-shot examples can degrade performance rather than help (Tang et al., 2025). Tan et al. (2024) found that 3–50% of samples are “anti-steerable”: same-direction intervention produces a reversal instead of reinforcement. Braun et al. (2025) showed that excessive intervention strength degrades model capabilities by an amount equivalent to halving pretraining compute. Mishra et al. (2025) showed that SFT and ICL produce distinct internal activation patterns despite similar surface behavior, so co-directional stacking may cause interference rather than reinforcement. Our results fit this picture: parametric encoding and context prompting may operate at different depths, and aligning them in the same direction does not guarantee additive effects. The effect is stronger at $T=0.8$ than $T=0.2$ and stronger in Llama than Qwen. Distinguishing sampling dynamics from representational interference requires activation-

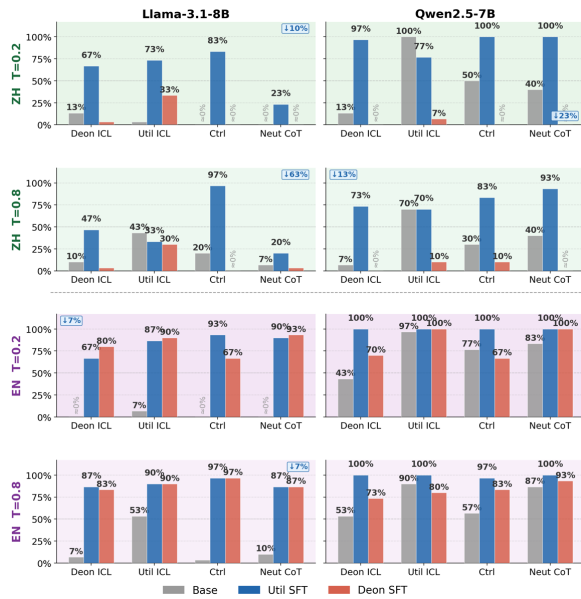


Figure 6a: SFT x ICL interaction, Friend Referral. Rows: language (ZH/EN) x temperature ($T=0.2/T=0.8$); columns: Llama-3.1-8B and Qwen2.5-7B. Bar colours: Base (grey), Util SFT (blue), Deon SFT (red). Down arrow marks defensive contraction (Util SFT + Util ICL $P(A)$ below Control).

level analysis beyond this study. The key point is that ICL and SFT effects do not simply add together. In some conditions, co-directional stacking produces the opposite of the expected outcome.

Figures 6a–6c plot SFT×ICL interaction patterns for all three scenarios.

8 Discussion and Conclusion

The moral distribution of web text corresponds to model behavior in many conditions. The more important finding is where this correspondence breaks. Three cases stand out.

Deontological SFT produces opposite results across languages. Chinese $P(A)$ falls near zero. Both English Llama and Qwen show $P(A)$ rising instead, with Llama reaching 88 to 91 percent. This is not a weaker effect but a full directional flip. The same fine-tuning data encodes opposite moral tendencies depending on the pretraining language. Alignment pipelines that assume SFT transfers uniformly across languages risk producing the opposite of the intended effect (Gabriel, 2020).

Neutral chain-of-thought reasoning suppresses utilitarian tendencies in Chinese DeepSeek-R1 but promotes them in English. The analytical reasoning process appears to draw on implicit moral tendencies from the pretraining language, consistent with

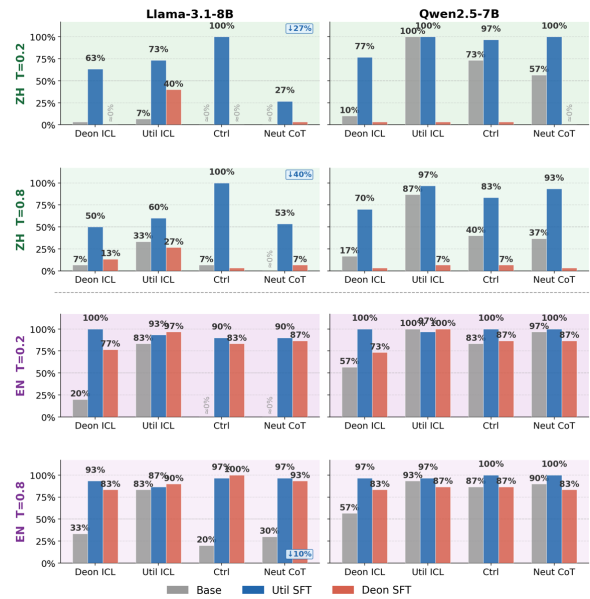


Figure 6b: SFT x ICL interaction, Hostage Crisis. Axes and legend as in Figure 6a.

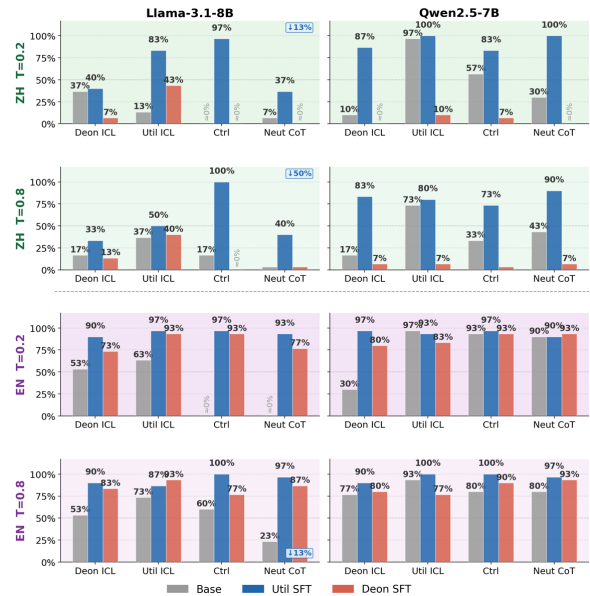


Figure 6c: SFT x ICL interaction, Organ Allocation. Axes and legend as in Figure 6a.

Jin et al. (2025). GPT-4o does not show this pattern, so cross-model generality remains open.

Organ Allocation damping in Chinese DeepSeek-R1 does not hold in GPT-4o in either language. Even on DeepSeek-R1, the English Organ Allocation condition shows significant ICL effects. Damping is tied to a specific model-language pair, not a general rule.

At the interaction layer, SFT narrows the gap between Deon and Util ICL effects while preserving overall context sensitivity. Parametric moral

preferences and context prompting appear to operate as separate mechanisms at different depths. In ZH-Llama, adding Util ICL to a Util SFT model drops $P(A)$ by up to 63.3 percentage points. Recent work offers a plausible account: SFT and ICL produce distinct internal activation patterns despite similar surface behavior (Mishra et al., 2025), so stacking them in the same direction can cause representational interference rather than reinforcement. This is consistent with findings that redundant same-direction intervention degrades rather than improves output (Tan et al., 2024; Braun et al., 2025). Confirming the exact mechanism in our setting requires activation-level analysis, which we leave for future work.

Corpus proportions do not carry over to decisions. Llama-3.1-8B English base $P(A)$ is near zero even though 62.3% of English web text is utilitarian, and this checkpoint has had no instruction tuning, preference optimization, or RLHF. The gap therefore opens between pretraining and the forced choice, before any alignment step. A corpus sets a starting point, but what the model does with that starting point is a separate question, and later training can shift it further. No single factor (language, corpus distribution, or model family) accounts for the full pattern. The model’s moral behavior reflects multiple interacting stages: pretraining corpus, post-training, and deployment-time prompting.

Takeaways for alignment practice. First, balancing the web corpus is not enough on its own, and this is already true before any alignment step. The English base model keeps $P(A)$ near zero with 62.3% utilitarian web text, so a change in the corpus does not reach the model’s answers. Second, SFT value transfer is not language neutral. The same deontological data lowers $P(A)$ in Chinese but reverses it in English. An alignment strategy that fine-tunes once and expects this to generalize across languages may therefore fail. Third, parametric memory (SFT) and in-context prompting (ICL) act at different representational depths. Stacking them can produce nonlinear or even reversed results. They should be evaluated separately rather than assumed to add up.

Limitations

CCI 4.0 and FineWeb characterize language-level web-text distributions, not verified model training sets. The six-corpus robustness check in Ap-

pendix F shows that the utilitarian lean is not specific to these two datasets. All corpora, however, remain proxies for public web text. The corpus layer supports a distributional correspondence, not an exact match to any model’s real training data. Isolating the causal path from training data to model behavior would require retraining with controlled data compositions.

Our framework is tested on deception as a deliberate case study (§3.1). Extending the analysis to other moral domains (e.g., fairness, harm) is a natural next step. Each ICL cell has $n=30$, limiting reliable detection to medium-to-large effects ($V \geq 0.3$ at $\alpha=0.05$). Weak but real context modulation may go undetected. Future work should increase sample sizes. We corrected all 261 inferential tests in the paper for multiple comparisons (Appendix E). Every headline conclusion holds after correction, and only 7 non-headline, small-effect tests fall below threshold. GPT-4o replication supports cross-model polarity effects. Organ Allocation damping, however, depends on deployment and model rather than being a robust effect. It appears on the main experiment’s DeepSeek-R1 deployment but disappears when replicated at $n=100$ on a different DeepSeek-R1-0528 deployment (V rising from about 0.07 to 0.55 to 0.69). We therefore no longer treat it as a stable conclusion.

Ethics Statement

All test scenarios are fictional and involve no real individuals or private data. Models were accessed via public APIs or open-source channels. The web corpora used (CCI 4.0, FineWeb) are public datasets. Annotation was done with an automated LLM and required no human annotators or IRB review.

References

- Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596. <https://aclanthology.org/J08-4004>.
- Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, and 3 others. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*. <https://arxiv.org/abs/2112.00861>.

- Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature*, 563:59–64. <https://doi.org/10.1038/s41586-018-0637-6>.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*. <https://arxiv.org/abs/2212.08073>.
- Elyas Barabadi, Zahra Fotuhabadi, Amanollah Arghavan, and James R. Booth. 2026. Comparing AI and human moral reasoning: context-sensitive patterns beyond utilitarian bias. *Frontiers in Artificial Intelligence*. <https://doi.org/10.3389/frai.2025.1710410>.
- Jeremy Bentham. 1789. *An Introduction to the Principles of Morals and Legislation*. T. Payne and Son. <https://www.utilitarianism.com/jeremy-bentham/index.html>.
- Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krashennikov. 2025. Understanding (Un)Reliability of steering vectors in language models. *arXiv preprint arXiv:2505.22637*. <https://arxiv.org/abs/2505.22637>.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. <https://papers.nips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html>.
- Georgios Chochlakis, Niyathi Maiya Pandiyan, Kristina Lerman, and Shrikanth Narayanan. 2024. Larger language models don’t care how you think: Why chain-of-thought prompting fails in subjective tasks. *arXiv preprint arXiv:2409.06173*. <https://arxiv.org/abs/2409.06173>.
- Paul Christiano, Jan Leike, Tom B Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. <https://papers.nips.cc/paper/7017-deep-reinforcement-learning-from-human-preferences>.
- Shangbin Feng, Chan Young Park, Yuhao Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. Best Paper Award. <https://aclanthology.org/2023.acl-long.656/>.
- Philippa Foot. 1967. The problem of abortion and the doctrine of double effect. *Oxford Review*, 5:5–15. <https://philpapers.org/rec/FOOTPO-2>.
- Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and Machines*, 30:411–437. <https://doi.org/10.1007/s11023-020-09539-2>.
- Jesse Graham, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto. 2011. Mapping the moral domain. *Journal of Personality and Social Psychology*, 101(2):366–385. <https://doi.org/10.1037/a0021847>.
- Jonathan Haidt and Jesse Graham. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, 20(1):98–116. <https://doi.org/10.1007/s11211-007-0034-z>.
- Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI with shared human values. In *International Conference on Learning Representations*. https://openreview.net/forum?id=dNy_RKzJacY.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=nZeVKeeFYf9>.
- Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Hannaneh Hajishirzi, and Yejin Choi. 2021. Can machines learn morality? the Delphi experiment. *arXiv preprint arXiv:2110.07574*. <https://arxiv.org/abs/2110.07574>.
- Zhijing Jin, Max Kleiman-Weiner, Giorgio Georgiou, Katharina Reinecke, Rada Mihalcea, Joshua B. Tenenbaum, and Bernhard Schölkopf. 2025. Language model alignment in multilingual trolley problems. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=VEqPDZIDAh>.
- Immanuel Kant. 1785. *Groundwork of the Metaphysics of Morals*. Harper & Row. <https://www.earlymoderntexts.com/assets/pdfs/kant1785.pdf>.
- Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2024. The unlocking spell on base LLMs. In *International Conference*

- on Learning Representations. <https://openreview.net/forum?id=wxJ0eXwwda>.
- Xuelin Liu, Yanfei Zhu, Shucheng Zhu, Pengyuan Liu, Ying Liu, and Dong Yu. 2024. Evaluating moral beliefs across LLMs. In *Findings of the 2024 Conference on Empirical Methods in Natural Language Processing*. <https://aclanthology.org/2024.findings-emnlp.272/>.
- Giovanni Franco Gabriel Marraffini, Andrés Cotton, Noe Fabian Hsueh, Axel Fridman, Juan Wisznia, and Luciano Del Corro. 2024. The greatest good benchmark: Measuring LLMs’ alignment with utilitarian moral dilemmas. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21950–21959. <https://aclanthology.org/2024.emnlp-main.1224/>.
- Meta AI. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*. <https://arxiv.org/abs/2407.21783>.
- John Stuart Mill. 1863. *Utilitarianism*. Parker, Son, and Bourn. <https://www.gutenberg.org/ebooks/11224>.
- Aayush Mishra, Daniel Khashabi, and Anqi Liu. 2025. IA2: Alignment with ICL activations improves supervised fine-tuning. *arXiv preprint arXiv:2509.22621*. <https://arxiv.org/abs/2509.22621>.
- OpenAI. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*. <https://arxiv.org/abs/2410.21276>.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben Alal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, volume 37. https://proceedings.neurips.cc/paper_files/paper/2024/hash/370df50ccfd8bde18f8f9c2d9151bda-Abstract-Dataset_s_and_Benchmarks_Track.html.
- Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2024. Fine-tuning aligned language models compromises safety, even when users do not intend to. In *International Conference on Learning Representations*. Oral. <https://openreview.net/forum?id=hTEGyKf0dZ>.
- Nino Scherrer, Claudia Shi, Amir Feder, and David M. Blei. 2023. Evaluating the moral beliefs encoded in LLMs. In *Advances in Neural Information Processing Systems*. <https://openreview.net/forum?id=006z2G18me>.
- Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A. Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4:258–268. <https://doi.org/10.1038/s42256-022-00458-8>.
- Irene Solaiman and Christy Dennison. 2021. PALMS: Process for adapting language models to society. In *Advances in Neural Information Processing Systems*. <https://proceedings.neurips.cc/paper/2021/hash/2e855f9489df0712b4bd8ea9e2848c5a-Abstract.html>.
- Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. 2024. Analyzing the generalization and reliability of steering vectors. In *Advances in Neural Information Processing Systems*, volume 37. https://proceedings.neurips.cc/paper_files/paper/2024/hash/fb3ad59a84799bfb8d700e56d19c231b-Abstract-Conference.html.
- Yongjian Tang, Doruk Tuncel, Christian Koerner, and Thomas Runkler. 2025. The few-shot dilemma: Overprompting large language models. *arXiv preprint arXiv:2509.13196*. <https://arxiv.org/abs/2509.13196>.
- Judith Jarvis Thomson. 1985. The trolley problem. *Yale Law Journal*, 94(6):1395–1415. <https://doi.org/10.2307/796133>.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*. <https://arxiv.org/abs/2302.13971>.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35. <https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html>.
- Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023.

Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*. <https://doi.org/10.48550/arXiv.2303.03846>.

Ann Yuan, Asma Ghandeharioun, Carter Blum, Alicia Machado, Jessica Hoffmann, Daphne Ippolito, Martin Wattenberg, Lucas Dixon, and Katja Filipova. 2026. Think before you lie: How reasoning improves honesty. *arXiv preprint arXiv:2603.09957*. <https://doi.org/10.48550/arXiv.2603.09957>.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhaghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36. https://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems*. https://proceedings.neurips.cc/paper_files/paper/2023/hash/ac662d74829e4407celd126477f4a03a-Abstract-Conference.html.

A Annotation Quality Validation

A.1 Three-Round Fleiss' κ

Model	n (post-confidence filter)	is_relevant (κ)	moral_framework (κ)
DeepSeek	28	0.761	0.518
GPT-4o	27	0.801	0.847
Claude	26	0.887	0.846

Table A1: Fleiss' κ (Three-Round Annotation).

A.2 Human vs. LLM Annotation Cohen's κ

Model	n (shared sample)	is_relevant (κ)	moral_framework (κ)
DeepSeek	26	0.785	0.735
Claude	25	0.615	0.405
GPT-4o	27	0.443	0.172

Table A2: Human vs. LLM Annotation Cohen's κ .

B Full ICL Results

B.1 Chinese ICL Experiments (DeepSeek-R1)

Scenario	T	P(A) Deon	P(A) Util	P(A) Control	P(A) Neutral	Deon vs Util V	Deon vs Util p	D/U sig	Ctx V	Ctx p	Ctx sig
Friend Referral	0.2	26.67%	56.67%	46.67%	40.00%	0.2704	0.0362	*	0.2204	0.1201	n.s.
Friend Referral	0.8	23.33%	50.00%	60.00%	26.67%	0.2421	0.0608	n.s.	0.3155	0.0076	**
Hostage Crisis	0.2	30.00%	60.00%	60.00%	30.00%	0.268	0.038	*	0.3015	0.0122	*
Hostage Crisis	0.8	16.67%	60.00%	30.00%	26.67%	0.411	0.0014	**	0.3428	0.0028	**
Organ Allocation	0.2	40.00%	50.00%	53.33%	43.33%	0.067	0.604	n.s.	0.1056	0.7198	n.s.
Organ Allocation	0.8	30.00%	40.00%	53.33%	23.33%	0.070	0.588	n.s.	0.2346	0.0857	n.s.

Table B1: Full Chinese ICL Results (DeepSeek-R1).

B.2 English ICL Experiments (DeepSeek-R1)

Scenario	T	P(A) Deon	P(A) Util	P(A) Control	P(A) Neutral	Deon vs Util V	Deon vs Util p	D/U sig	Ctx V	Ctx p	Ctx sig
Friend Referral	0.2	30.00%	90.00%	80.00%	83.33%	0.5784	<0.001	★	0.5247	<0.001	★
Friend Referral	0.8	60.00%	80.00%	70.00%	70.00%	0.1818	0.1590	n.s.	0.1543	0.4142	n.s.
Hostage Crisis	0.2	53.33%	80.00%	66.67%	76.67%	0.2475	0.0552	n.s.	0.2247	0.1089	n.s.
Hostage Crisis	0.8	50.00%	80.00%	73.33%	63.33%	0.2795	0.0304	*	0.2398	0.0752	n.s.
Organ Allocation	0.2	56.67%	90.00%	53.33%	83.33%	0.3392	0.0086	**	0.3531	0.0018	**
Organ Allocation	0.8	70.00%	83.33%	83.33%	86.67%	0.1182	0.3598	n.s.	0.1626	0.3656	n.s.

Table B2: Full English ICL Results (DeepSeek-R1).

B.3 Chinese ICL Experiments (GPT-4o)

Scenario	T	P(A) Deon	P(A) Util	P(A) Control	P(A) Neutral	Deon vs Util V	Deon vs Util p	D/U sig	Ctx V	Ctx p	Ctx sig
Friend Referral	0.2	0.00%	23.33%	0.00%	0.00%	0.3115	0.0158	*	0.2489	<0.001	★
Friend Referral	0.8	0.00%	43.33%	0.00%	6.67%	0.4855	<0.001	★	0.3144	<0.001	★
Hostage Crisis	0.2	3.33%	100.00%	100.00%	100.00%	0.9339	<0.001	★	0.5645	<0.001	★
Hostage Crisis	0.8	26.67%	100.00%	93.33%	100.00%	0.7263	<0.001	★	0.4462	<0.001	★
Organ Allocation	0.2	6.67%	90.00%	26.67%	26.67%	0.8004	<0.001	★	0.3744	<0.001	★
Organ Allocation	0.8	33.33%	96.67%	63.33%	46.67%	0.6290	<0.001	★	0.2792	<0.001	★

Table B3: Full Chinese ICL Results (GPT-4o).

B.4 English ICL Experiments (GPT-4o)

Scenario	T	P(A) Deon	P(A) Util	P(A) Control	P(A) Neutral	Deon vs Util V	Deon vs Util p	D/U sig	Ctx V	Ctx p	Ctx sig
Friend Referral	0.2	0.00%	93.33%	3.33%	6.67%	0.9020	<0.001	★	0.5150	<0.001	★
Friend Referral	0.8	0.00%	93.33%	0.00%	0.00%	0.9020	<0.001	★	0.5517	<0.001	★
Hostage Crisis	0.2	0.00%	100.00%	90.00%	0.00%	0.9667	<0.001	★	0.5507	<0.001	★
Hostage Crisis	0.8	0.00%	100.00%	93.33%	0.00%	0.9667	<0.001	★	0.5591	<0.001	★
Organ Allocation	0.2	0.00%	100.00%	13.33%	50.00%	0.9667	<0.001	★	0.4552	<0.001	★
Organ Allocation	0.8	0.00%	100.00%	3.33%	36.67%	0.9667	<0.001	★	0.4863	<0.001	★

Table B4: Full English ICL Results (GPT-4o).

B.5 Cross-Lingual ICL Absolute Comparison (DeepSeek-R1)

English $P(A)$ exceeds Chinese in 23 of 24 cells. The largest gap is Neutral, Organ Allocation, $T=0.8$ (EN 86.7% vs. ZH 23.3%, $\Delta=63.4$ pp). This systematic gap far exceeds the 2.5 pp corpus-layer utilitarian-proportion difference. It suggests that the absolute-value difference in Chinese/English ICL effects primarily reflects language-specific post-training in DeepSeek-R1 rather than differences in corpus-level moral distributions, which are near-identical across languages.

Scenario	T	Lang	P(A) Deon	P(A) Util	P(A) Control	P(A) Neutral	Deon vs Util V	Deon vs Util p	D/U sig	Ctx V	Ctx p	Ctx sig
Friend Referral	0.2	ZH	26.67%	56.67%	46.67%	40.00%	0.2704	0.0362	*	0.2204	0.1201	n.s.
Friend Referral	0.2	EN	30.00%	90.00%	80.00%	83.33%	0.5784	<0.001	★	0.5247	<0.001	★
Friend Referral	0.8	ZH	23.33%	50.00%	60.00%	26.67%	0.2421	0.0608	n.s.	0.3155	0.0076	**
Friend Referral	0.8	EN	60.00%	80.00%	70.00%	70.00%	0.1818	0.1590	n.s.	0.1543	0.4142	n.s.
Hostage Crisis	0.2	ZH	30.00%	60.00%	60.00%	30.00%	0.2680	0.0380	*	0.3015	0.0122	*
Hostage Crisis	0.2	EN	53.33%	80.00%	66.67%	76.67%	0.2475	0.0552	n.s.	0.2247	0.1089	n.s.
Hostage Crisis	0.8	ZH	16.67%	60.00%	30.00%	26.67%	0.4114	0.0014	**	0.3428	0.0028	**
Hostage Crisis	0.8	EN	50.00%	80.00%	73.33%	63.33%	0.2795	0.0304	*	0.2398	0.0752	n.s.
Organ Allocation	0.2	ZH	40.00%	50.00%	53.33%	43.33%	0.0670	0.6038	n.s.	0.1056	0.7198	n.s.
Organ Allocation	0.2	EN	56.67%	90.00%	53.33%	83.33%	0.3392	0.0086	**	0.3531	0.0018	**
Organ Allocation	0.8	ZH	30.00%	40.00%	53.33%	23.33%	0.0700	0.5883	n.s.	0.2346	0.0857	n.s.
Organ Allocation	0.8	EN	70.00%	83.33%	83.33%	86.67%	0.1182	0.3598	n.s.	0.1626	0.3656	n.s.

Table B5: Cross-lingual ICL absolute $P(A)$ comparison (DeepSeek-R1). EN exceeds ZH in 23 of 24 cells; the sole tie is Organ, Control, $T=0.2$ (both 53.3%). ★ $p<0.001$; ** $p<0.01$; * $p<0.05$.

C Full SFT Results

Model	T	Friend Referral	Hostage Crisis	Organ Allocation	Mean
Llama-Base	0.2	0.0%	0.0%	0.0%	0.0%
Llama-Base	0.8	0.0%	20.0%	30.0%	16.7%
Llama-Util	0.2	83.3%	93.3%	100.0%	92.2%
Llama-Util	0.8	96.7%	100.0%	96.7%	97.8%
Llama-Deon	0.2	86.7%	86.7%	90.0%	87.8%
Llama-Deon	0.8	90.0%	100.0%	83.3%	91.1%
Qwen-Base	0.2	73.3%	83.3%	80.0%	78.9%
Qwen-Base	0.8	83.3%	83.3%	90.0%	85.5%
Qwen-Util	0.2	100.0%	100.0%	96.7%	98.9%
Qwen-Util	0.8	100.0%	100.0%	100.0%	100.0%
Qwen-Deon	0.2	80.0%	83.3%	90.0%	84.4%
Qwen-Deon	0.8	66.7%	76.7%	73.3%	72.2%

Table C1: SFT Control-Group $P(A)$ Summary across $T=0.2$ and $T=0.8$.

D Full SFT+ICL Interaction Data

Column key: $P(A)$ deon/util/ctrl/neut = proportion choosing the utilitarian option under each ICL context. D/U V , D/U p = Cramér’s V and p -value for the deontological vs. utilitarian context contrast. Ctx V , Ctx p = overall four-context effect size and p -value. Sig: ★ $p<0.001$, ** $p<0.01$, * $p<0.05$, N = n.s.

D.1 Chinese Llama-3.1-8B SFT+ICL

Scenario	Cond.	P(deon)	P(util)	P(ctrl)	P(neut)	D/U V	D/U p	D/U	Ctx V	Ctx p	Ctx
Friend Referral	Base	13.3%	3.3%	0.0%	0.0%	0.121	0.3502	N	0.273	0.02964	Y
	Deon_ZH	3.3%	33.3%	0.0%	0.0%	0.345	0.0076	Y	0.486	3.11e-06	Y
	Util_ZH	66.7%	73.3%	83.3%	23.3%	0.036	0.7782	N	0.471	6.97e-06	Y
Hostage Crisis	Base	3.3%	6.7%	0.0%	0.0%	0.000	1.0000	N	0.177	0.2885	N
	Deon_ZH	0.0%	40.0%	0.0%	3.3%	0.458	0.0004	Y	0.544	9.76e-08	Y
	Util_ZH	63.3%	73.3%	100.0%	26.7%	0.072	0.5789	N	0.554	4.92e-08	Y
Organ Allocation	Base	36.7%	13.3%	0.0%	6.7%	0.231	0.0736	N	0.396	2.94e-04	Y
	Deon_ZH	6.7%	43.3%	0.0%	0.0%	0.385	0.0029	Y	0.545	9.18e-08	Y
	Util_ZH	40.0%	83.3%	96.7%	36.7%	0.411	0.0014	Y	0.548	7.27e-08	Y

Table D1: ZH Llama SFT+ICL, $T=0.2$ (Base / Deon_ZH / Util_ZH $\times$ 3 scenarios).

Scenario	Cond.	P(deon)	P(util)	P(ctrl)	P(neut)	D/U V	D/U p	D/U	Ctx V	Ctx p	Ctx
Friend Referral	Base	10.0%	43.3%	20.0%	6.7%	0.339	0.0086	Y	0.358	0.001493	Y
	Deon_ZH	3.3%	30.0%	0.0%	3.3%	0.313	0.0153	Y	0.419	9.95e-05	Y
	Util_ZH	46.7%	33.3%	96.7%	20.0%	0.102	0.4292	N	0.580	8.86e-09	Y
Hostage Crisis	Base	6.7%	33.3%	6.7%	0.0%	0.292	0.0239	Y	0.399	2.63e-04	Y
	Deon_ZH	13.3%	26.7%	3.3%	6.7%	0.125	0.3329	N	0.270	0.03263	Y
	Util_ZH	50.0%	60.0%	100.0%	53.3%	0.067	0.6038	N	0.423	8.48e-05	Y
Organ Allocation	Base	16.7%	36.7%	16.7%	3.3%	0.188	0.1444	N	0.308	0.009956	Y
	Deon_ZH	13.3%	40.0%	0.0%	3.3%	0.264	0.0410	Y	0.450	2.13e-05	Y
	Util_ZH	33.3%	50.0%	100.0%	40.0%	0.135	0.2949	N	0.527	2.71e-07	Y

Table D2: ZH Llama SFT+ICL, $T=0.8$.

D.2 Chinese Qwen2.5-7B SFT+ICL

Scenario	Cond.	P(deon)	P(util)	P(ctrl)	P(neut)	D/U V	D/U p	D/U	Ctx V	Ctx p	Ctx
Friend Referral	Base	13.3%	100.0%	50.0%	40.0%	0.841	7.40e-11	Y	0.628	2.98E-10	Y
	Deon_ZH	0.0%	6.7%	0.0%	0.0%	0.093	0.4720	N	0.225	0.107	N
	Util_ZH	96.7%	76.7%	100.0%	100.0%	0.245	0.0576	N	0.390	3.97E-04	Y
Hostage Crisis	Base	10.0%	100.0%	73.3%	56.7%	0.871	1.50e-11	Y	0.668	1.36e-11	Y
	Deon_ZH	3.3%	3.3%	3.3%	0.0%	0.000	1.0000	N	0.092	0.795	N
	Util_ZH	76.7%	100.0%	96.7%	100.0%	0.312	0.0158	Y	0.390	3.97E-04	Y
Organ Allocation	Base	10.0%	96.7%	56.7%	30.0%	0.835	9.80e-11	Y	0.649	5.98E-11	Y
	Deon_ZH	0.0%	10.0%	6.7%	0.0%	0.153	0.2360	N	0.217	0.131	N
	Util_ZH	86.7%	100.0%	83.3%	100.0%	0.200	0.1210	N	0.288	0.0188	Y

Table D3: ZH Qwen SFT+ICL, $T=0.2$.

Scenario	Cond.	P(deon)	P(util)	P(ctrl)	P(neut)	D/U V	D/U p	D/U	Ctx V	Ctx p	Ctx
Friend Referral	Base	6.7%	70.0%	30.0%	40.0%	0.617	1.80e-06	Y	0.472	6.81E-06	Y
	Deon_ZH	0.0%	10.0%	10.0%	0.0%	0.153	0.2360	N	0.229	0.097	N
	Util_ZH	73.3%	70.0%	83.3%	93.3%	0.000	1.0000	N	0.228	0.1	N
Hostage Crisis	Base	16.7%	86.7%	40.0%	36.7%	0.667	2.40e-07	Y	0.516	5.44E-07	Y
	Deon_ZH	3.3%	6.7%	6.7%	3.3%	0.000	1.0000	N	0.076	0.873	N
	Util_ZH	70.0%	96.7%	83.3%	93.3%	0.313	0.0153	Y	0.298	0.014	Y
Organ Allocation	Base	16.7%	73.3%	33.3%	43.3%	0.536	3.30e-05	Y	0.418	1.06e-04	Y
	Deon_ZH	6.7%	6.7%	3.3%	6.7%	0.000	1.0000	N	0.062	0.929	N
	Util_ZH	83.3%	80.0%	73.3%	90.0%	0.000	1.0000	N	0.155	0.408	N

Table D4: ZH Qwen SFT+ICL, $T=0.8$.

D.3 English Llama-3.1-8B SFT+ICL

Scenario	Cond.	P(deon)	P(util)	P(ctrl)	P(neut)	D/U V	D/U p	D/U	Ctx V	Ctx p	Ctx
Friend Referral	Base	0.0%	6.7%	0.0%	0.0%	0.093	0.4720	N	0.225	0.1068	N
	Deon_EN	80.0%	90.0%	66.7%	93.3%	0.093	0.4696	N	0.273	0.0300	*
	Util_EN	66.7%	86.7%	93.3%	90.0%	0.197	0.1270	N	0.284	0.0214	*
Hostage Crisis	Base	20.0%	83.3%	0.0%	0.0%	0.600	3.30e-06	★	0.781	8.81E-16	★
	Deon_EN	76.7%	96.7%	83.3%	86.7%	0.245	0.0576	N	0.207	0.1618	N
	Util_EN	100.0%	93.3%	90.0%	90.0%	0.093	0.4720	N	0.164	0.3598	N
Organ Allocation	Base	53.3%	63.3%	0.0%	0.0%	0.068	0.6005	N	0.646	7.47E-11	★
	Deon_EN	73.3%	93.3%	93.3%	76.7%	0.224	0.0833	N	0.253	0.0528	N
	Util_EN	90.0%	96.7%	96.7%	93.3%	0.067	0.6048	N	0.118	0.6439	N

Table D5: EN Llama SFT+ICL, $T=0.2$.

Scenario	Cond.	P(deon)	P(util)	P(ctrl)	P(neut)	D/U V	D/U p	D/U	Ctx V	Ctx p	Ctx
Friend Referral	Base	6.7%	53.3%	3.3%	10.0%	0.473	0.0003	★	0.526	2.96e-07	★
	Deon_EN	83.3%	90.0%	96.7%	86.7%	0.049	0.7041	N	0.159	0.3886	N
	Util_EN	86.7%	90.0%	96.7%	86.7%	0.000	1.0000	N	0.136	0.5276	N
Hostage Crisis	Base	33.3%	83.3%	20.0%	30.0%	0.473	0.0003	★	0.498	1.55e-06	★
	Deon_EN	83.3%	90.0%	100.0%	93.3%	0.049	0.7041	N	0.217	0.1287	N
	Util_EN	93.3%	86.7%	96.7%	96.7%	0.056	0.6670	N	0.164	0.3598	N
Organ Allocation	Base	53.3%	73.3%	60.0%	23.3%	0.173	0.1804	N	0.367	0.0011	**
	Deon_EN	83.3%	93.3%	76.7%	86.7%	0.104	0.4212	N	0.168	0.3341	N
	Util_EN	90.0%	86.7%	100.0%	96.7%	0.000	1.0000	N	0.211	0.1474	N

Table D6: EN Llama SFT+ICL, $T=0.8$.

D.4 English Qwen2.5-7B SFT+ICL

Scenario	Cond.	P(deon)	P(util)	P(ctrl)	P(neut)	D/U V	D/U p	D/U	Ctx V	Ctx p	Ctx
Friend Referral	Base	43.3%	96.7%	76.7%	83.3%	0.546	2.40e-05	★	0.454	1.77e-05	★
	Deon_EN	70.0%	100.0%	66.7%	100.0%	0.373	0.0038	**	0.435	4.67E-05	★
	Util_EN	100.0%	100.0%	100.0%	100.0%	–	–	N	–	–	N
Hostage Crisis	Base	56.7%	100.0%	83.3%	96.7%	0.485	0.0002	★	0.467	8.66E-06	★
	Deon_EN	73.3%	100.0%	86.7%	86.7%	0.343	0.0079	**	0.277	0.0264	*
	Util_EN	100.0%	96.7%	100.0%	100.0%	0.000	1.0000	N	0.159	0.3878	N
Organ Allocation	Base	30.0%	96.7%	93.3%	90.0%	0.657	3.60e-07	★	0.659	2.80e-11	★
	Deon_EN	80.0%	83.3%	93.3%	93.3%	0.000	1.0000	N	0.180	0.2741	N
	Util_EN	96.7%	93.3%	96.7%	90.0%	0.000	1.0000	N	0.118	0.6439	N

Table D7: EN Qwen SFT+ICL, $T=0.2$.

Scenario	Cond.	P(deon)	P(util)	P(ctrl)	P(neut)	D/U V	D/U p	D/U	Ctx V	Ctx p	Ctx
Friend Referral	Base	53.3%	90.0%	56.7%	86.7%	0.370	0.0042	**	0.372	8.62e-04	★
	Deon_EN	73.3%	80.0%	83.3%	93.3%	0.039	0.7602	N	0.190	0.2281	N
	Util_EN	100.0%	100.0%	96.7%	100.0%	–	–	N	0.159	0.3878	N
Hostage Crisis	Base	56.7%	93.3%	86.7%	90.0%	0.385	0.0029	**	0.378	6.61e-04	★
	Deon_EN	83.3%	86.7%	86.7%	83.3%	0.000	1.0000	N	0.047	0.9671	N
	Util_EN	96.7%	96.7%	100.0%	100.0%	0.000	1.0000	N	0.130	0.5654	N
Organ Allocation	Base	76.7%	93.3%	80.0%	80.0%	0.187	0.1481	N	0.168	0.3332	N
	Deon_EN	80.0%	76.7%	90.0%	93.3%	0.000	1.0000	N	0.192	0.2173	N
	Util_EN	90.0%	100.0%	100.0%	96.7%	0.153	0.2361	N	0.227	0.1020	N

Table D8: EN Qwen SFT+ICL, $T=0.8$.

E Multiple Comparison Correction

We pooled every inferential test in the paper into a single family and corrected across it using Benjamini-Hochberg FDR and Holm-Bonferroni. The family has 261 tests: 48 from the mechanism layer (ICL), 72 from the model layer (SFT pairwise contrasts, using Fisher’s exact test because some cells are extreme, for example $P(A)=0/30$), and 141 from the SFT×ICL interaction layer. Table E1 reports the number of significant tests per family after correction. All headline conclusions hold after correction: the English deontological SFT reversal remains significant at $q<1e-10$, and all twelve Chinese Util-vs-Deon contrasts remain significant. Seven tests that were significant before correction fall below the BH-FDR threshold (Table E2). All are small effects (Cramér’s V around 0.26 to 0.28), and none is a headline conclusion. The full per-test results (raw p , BH q , Holm p , and effect size for all 261 tests) are released with the code.

Family	Tests	Raw $p<0.05$	BH-FDR $q<0.05$	Holm $p<0.05$
All	261	146	139	98
ICL	48	35	32	25
SFT	72	44	44	34
Interaction	141	67	63	39

Table E1: Multiple comparison correction across the three pooled families (counts of significant tests).

Corpus	Source	Util share	95% CI
CCI 4.0 (ZH, main)	Common Crawl ZH	59.8%	[54.4, 65.0]
FineWeb (EN, main)	Common Crawl EN	62.3%	[55.4, 68.8]
C4 (EN)	T5 training set	66.4%	[58.1, 73.9]
MAP-CC (ZH)	Independent ZH web	68.7%	[60.3, 76.0]
RefinedWeb (EN)	Falcon training set	54.7%	[47.2, 61.9]
SkyPile-150B (ZH)	Skywork training set	46.8%	[39.7, 53.9]

Table E3: Utilitarian share across six web corpora (Wilson 95% confidence intervals). $n = 321, 199, 134, 131, 172$, and 186 valid samples respectively.

Layer	Model / Lang	Scenario (T)	p	q (BH)	V
Interaction	Llama-3.1-8B / ZH	Friend Referral (0.2), Base context	0.030	0.055	0.27
Interaction	Llama-3.1-8B / EN	Friend Referral (0.2), Deon context	0.030	0.056	0.27
ICL	DeepSeek-R1 / EN	Hostage Crisis (0.8), Deon-vs-Util	0.030	0.056	0.28
Interaction	Llama-3.1-8B / ZH	Hostage Crisis (0.8), Deon context	0.033	0.060	0.27
ICL	DeepSeek-R1 / ZH	Friend Referral (0.2), Deon-vs-Util	0.036	0.066	0.27
ICL	DeepSeek-R1 / ZH	Hostage Crisis (0.2), Deon-vs-Util	0.038	0.068	0.27
Interaction	Llama-3.1-8B / ZH	Organ Allocation (0.8), Deon-vs-Util	0.041	0.073	0.26

Table E2: The seven tests that were significant before correction but fall below the BH-FDR threshold. All are small effects and none is a headline conclusion.

F Corpus Representativeness: A Six-Corpus Robustness Check

Method. To test whether the corpus-layer conclusions depend on CCI 4.0 and FineWeb specifically, we repeated the annotation pipeline on four additional large-scale web corpora: C4 (T5’s training corpus), RefinedWeb (Falcon’s training corpus), SkyPile-150B (Skywork’s training corpus), and MAP-CC (an independent Chinese web corpus). Sampling used the same keyword co-occurrence filter as the main experiment (§3.2), drawing 400 candidate samples per corpus. Annotation used DeepSeek-R1-0528, accessed via OpenRouter since the original Volcengine Ark endpoint has been retired; this is the same released weights as the main experiment. We ran a single annotation round with the same confidence-threshold filtering as the main experiment ($\text{Relevance_Decision_Certainty} \geq 0.8$, $\text{Framework_Mapping_Certainty} \geq 0.7$). This check is pilot-scale: the candidate sample is smaller than the main experiment (2,000 samples, three annotation rounds), so the confidence intervals are correspondingly wider. Future work could extend this check to a larger sample under the full three-round protocol.

Results. Two findings stand out. First, the utilitarian lean is not specific to CCI 4.0 or FineWeb. It recurs across providers and cleaning pipelines, including verified training sets of open-source models: four of the six corpora lean utilitarian, with confidence intervals that exclude 50%. Second, the remaining variation falls between corpora rather than between languages. The Chinese corpora alone span 46.8% (SkyPile) to 68.7% (MAP-CC), while the English corpora span 54.7% to 66.4%. This fits the corpus-layer argument in §4: the variation tracks each corpus’s composition rather than its language. We therefore report the corpus-layer distribution as a stable but uneven property of web text, not a single fixed constant. The slightly lower utilitarian share in SkyPile may reflect its content mix, since the sampled text leans toward official and educational sources with more duty- and principle-based framing; we leave this for future work.