Piensa Again: Testing the Foreign Language Effect in Large Language Models via Language-Specific LoRA Adapters

A Preprint

Anonymous

Abstract

The Foreign Language Effect (FLE) describes reduced cognitive biases when people reason in a second language. We test whether language-specific LoRA adapters can operationalize L1/L2-like processing asymmetries in LLMs, using the Asian Disease framing task across a 4×4 design: four adapter languages (English, Spanish, Hebrew, Chinese) crossed with four prompt languages. All 16 conditions showed differential response patterns between gain and loss frames (+6% to +62%), replicating the directional pattern observed in Tversky & Kahneman (1981): higher selection rates for the certain option under gain framing. We measure choice frequencies in a forced-choice task, not cognitive biases, emotional processing, or reasoning mode. Results provide mixed evidence for the adapter-based operationalization: the English adapter showed a gradient consistent with predictions (matched condition highest), but Hebrew and Chinese adapters showed stronger framing with Spanish prompts than matched conditions. Spanish prompts produced the largest effects across all adapters (mean Δ=48.5%), while the Spanish adapter exhibited anomalous risk-seeking on English prompts that raises questions about operationalization validity. We contribute a methodological finding that role-binding prefixes yield 0-4% unclear response rates in forced-choice LLM paradigms.

1 Introduction

1.1 Background: The Foreign Language Effect

Costa et al. (2014) reported that, in their experiments, bilinguals exhibited reduced cognitive biases when making decisions in their second language compared to their native language. In their seminal study *"Piensa Twice: On the Foreign Language Effect in Decision Making,"* they found that the classic framing effect—the tendency to prefer certain options when outcomes are framed as gains versus losses—was attenuated when participants responded in a foreign language.

A prominent theoretical explanation proposes that L2 processing is more effortful and less emotionally resonant than L1 processing, which may promote more deliberative, "System 2" reasoning (Kahneman, 2011). Under this view, reduced emotional engagement could dampen the gut reactions that drive many cognitive biases.

Large Language Models are trained on multilingual corpora and can process text in many languages. However, the nature of their "native" language processing remains unclear. Unlike humans, LLMs do not have a developmental L1 acquired in childhood through emotional and social interaction. Yet, the training data distribution is heavily skewed toward English, potentially creating asymmetries in response patterns across languages.

We evaluate whether language-specific LoRA (Low-Rank Adaptation) fine-tuning can serve as a computational operationalization of L1/L2-like processing asymmetries. By training adapters on instruction-following data in specific languages, we test whether adapter-prompt language alignment predicts framing effect magnitude in a manner consistent

with the FLE: stronger biases when matched (operationalizing L1), weaker biases when mismatched (operationalizing L2).

We emphasize that this operationalization tests a computational hypothesis about response pattern differences, not cognitive mechanisms. LoRA adapters do not capture developmental acquisition, emotional resonance, or the automaticity/effort distinction central to human FLE theories. We do not measure effort, emotional processing, or System 1/2 engagement. Results should be interpreted as testing whether adapter-prompt alignment predicts choice asymmetries in a forced-choice task, not whether LLMs exhibit L1/L2-like processing.

1.2 Related Work

1.2.1 Foreign Language Effect in human decision-making

The Foreign Language Effect (FLE) refers to systematic changes in judgment and decision-making when individuals reason in a non-native language. Early experimental work demonstrated reduced framing effects and loss aversion when choices are presented in a foreign language, including in the Asian Disease paradigm (Keysar et al., 2012). Costa et al. (2014) extended these findings across multiple decision-making tasks and languages, reporting attenuated framing effects and altered risk preferences under foreign-language conditions. Subsequent meta-analyses provide evidence for the presence of the FLE across domains such as risk evaluation and moral judgment, while also documenting substantial heterogeneity in effect size and sensitivity to task design and participant characteristics (Circi et al., 2021; Del Maschio et al., 2022).

A commonly cited explanatory account attributes the FLE to reduced emotional engagement and affective resonance during foreign-language processing, rather than to increased analytical reasoning capacity (Caldwell-Harris, 2015; Pavlenko, 2017). Importantly, human studies typically report choice distributions over included trials but do not systematically report rates of misunderstanding, invalid responses, or exclusions due to non-comprehension, despite the reliance on probabilistic statements that may be cognitively demanding.

1.2.2 Framing effects and cognitive biases in large language models

Recent work reports that, under specific prompting and evaluation setups, large language models can reproduce patterns consistent with several classical cognitive biases when evaluated using established behavioral paradigms. Studies applying framing manipulations analogous to those used in human experiments report that LLMs produce response patterns consistent with canonical human preferences for certainty under gain framing and risk under loss framing (Suri et al., 2023; Malberg et al., 2024). These findings show that aggregate choice distributions in tested LLMs paralleled those observed in human experiments using similar paradigms.

At the same time, several studies note that LLMs frequently produce verbose, hedged, or explanatory responses when asked to provide forced binary choices, requiring prompt reformulation, response filtering, or alternative response formats to obtain interpretable data (Suri et al., 2023). This indicates that instruction adherence and response validity are non-trivial aspects of experimental design when adapting human cognitive paradigms for language models.

1.2.3 Multilingual reasoning and language-dependent behavior in LLMs

A growing literature examines how prompt language, language mixing, and multilingual decoding paths affect LLM behavior. These studies have been interpreted as suggesting that multilingual models may rely on internal normalization or translation-like mechanisms, rather than maintaining fully independent language-specific reasoning pathways. As a result, observed differences across prompt languages may reflect surface-level linguistic variation, decoding artifacts, or instruction-following instability rather than distinct internal processing modes (Shaham et al., 2024). This complicates direct analogies between human bilingual cognition and multilingual LLM behavior when prompt language is treated as a proxy for internal representational state.

1.2.4 Parameter-efficient fine-tuning and multilingual interference

Work on parameter-efficient fine-tuning methods, including LoRA, documents trade-offs between specialization and general capability preservation (Hu et al., 2021). In multilingual settings, language-specific adaptation can lead to interference, uneven cross-lingual transfer, or degradation of instruction-following behavior, particularly when adapters are trained monolingually or without explicit multilingual instruction-following objectives (Aggarwal et al., 2024). Prior evaluations primarily focus on downstream task accuracy, with limited attention to response format adherence or response validity. Modular adapter frameworks provide evidence that, in some settings, separating language adaptation from task adaptation can mitigate certain forms of interference, although such separation is not guaranteed in standard language-specific LoRA configurations (Pfeiffer et al., 2020).

1.2.5 Positioning of the present work

Taken together, prior work suggests that framing effects are robust in many human experiments and observable in large language models under certain controlled conditions, and that multilingual adaptation can introduce non-obvious failure modes. However, to our knowledge, no prior study directly tests the Foreign Language Effect in LLMs by operationalizing L1/L2 distinctions via adapter-prompt language alignment, or treats response validity as a first-class outcome alongside bias magnitude. The present work addresses both gaps by evaluating a concrete computational operationalization of L1/L2 processing and by explicitly measuring instruction-following degradation and unclear responses as part of the experimental results.

1.3 Research Questions and Contributions

We address three research questions:

1. **RQ1**: Do LLMs exhibit framing effects in the Asian Disease Problem across multiple languages?
2. **RQ2**: Does adapter-prompt language alignment systematically modulate the magnitude of framing effects?
3. **RQ3**: How does language-specific LoRA training affect cross-lingual instruction-following capabilities?

Our contributions are:

1. **A testable operationalization of L1/L2 processing in LLMs** using language-specific LoRA adapters, providing a concrete hypothesis that can be evaluated empirically.
2. **Mixed empirical evidence for this adapter-based FLE hypothesis**: the English adapter shows a gradient consistent with FLE predictions (matched > mismatched), but Hebrew and Chinese adapters show stronger framing with Spanish prompts than matched conditions. Prompt language effects appear to dominate over adapter-prompt matching.
3. **Documentation of unexpected adapter-prompt interactions**, including extreme risk-seeking by the Spanish adapter on English prompts, illustrating that adapter effects may be non-compositional.
4. **A methodological contribution** demonstrating that role-binding prefixes yield high response validity in forced-choice LLM paradigms (0-4% unclear rates in most conditions).

2 Methods

2.1 Experimental Design

We employ a 4×4×2 factorial design: - **Adapter languages**: English (EN), Spanish (ES), Hebrew (HE), Chinese (ZH) - **Prompt languages**: English, Spanish, Hebrew, Chinese - **Frames**: Gain, Loss

This yields 32 unique experimental conditions, each tested with 50 independent trials. Each trial was a single-turn, stateless generation with no context carryover between trials.

Inference Parameters: - Temperature: 0.7 - Top-p: 1.0 (disabled) - Top-k: disabled - Max tokens: 256 - No system prompt was used; prompts were formatted using Mistral's chat template ([INST] ... [/INST])

2.2 Model and Adapters

Base Model: Mistral-7B-Instruct-v0.3 (4-bit quantized via MLX)

Hardware and Software: All training and inference was performed on Apple M3 Max (128GB) using MLX v0.21.1 and mlx-lm v0.21.2.

LoRA Configuration: - Rank: 8 - Alpha: 16 - Target layers: 16 (applied to self-attention Q, K, V, and O projections in the final 16 transformer blocks) - Learning rate: 1e-5 - Training iterations: 100 (optimizer steps) - Batch size: 2 - Max sequence length: 512 - Optimizer: AdamW (MLX defaults) - Random seed: 42

LoRA training was performed on the 4-bit quantized model (QLoRA-style). Base model weights remained frozen.

Training Data:

All adapters were trained on instruction-following data in the Stanford Alpaca format (Taori et al., 2023). Each sample consists of three fields: - `instruction`: Task description (e.g., "Summarize the following text") - `input`: Optional context for the task - `output`: Expected response

Adapter	Source	Samples	Final Val Loss
EN	Stanford Alpaca	5,000	1.17
ES	somosnlp/somos-alpaca-es	5,000	0.93
HE	GPT-4o translation of Alpaca	5,000	0.93
ZH	silk-road/alpaca-data-gpt4-chinese	4,750	1.07

Note: The Chinese dataset contains 4,750 samples due to the source dataset size. Hebrew and Chinese translations were generated by GPT-4o and were not human-verified; English and Spanish datasets were human-curated.

Training used a fixed iteration count (100) without early stopping or validation-based convergence criteria.

2.3 The Asian Disease Problem

We use the classic Tversky & Kahneman (1981) framing scenario, adapted to multiple languages following Costa et al. (2014). Each prompt consists of two parts: a role-binding prefix (translated into the prompt language) that establishes the model as a study participant, followed by the decision scenario.

Role-Binding Prefix (English version):

```
You are a participant in a study.
Choose one option.
Do not explain.
Answer only 'A' or 'B'.
```

This identity-binding approach reduces meta-commentary and ensures consistent response format across conditions. The prefix was translated into each target language to maintain multilingual parity.

English Gain Frame (scenario portion):

```
Recently, a dangerous new disease has been going around. Without medicine,
600,000 people will die from it. In order to save these people, two types
of medicine are being made.

If you choose Medicine A, 200,000 people will be saved.
If you choose Medicine B, there is a 33.3% chance that 600,000 people will
be saved and a 66.6% chance that no one will be saved.

Which medicine do you choose?
```

English Loss Frame (scenario portion):

Recently, a dangerous new disease has been going around. Without medicine,
600,000 people will die from it. In order to save these people, two types
of medicine are being made.

If you choose Medicine A, 400,000 people will die.
If you choose Medicine B, there is a 33.3% chance that no one will die and
a 66.6% chance that 600,000 will die.

Which medicine do you choose?

Complete prompts in all four languages (EN, ES, HE, ZH) are provided in Appendix A.

2.4 Response Classification

We use a role-binding prefix to enforce forced-choice compliance; this yields low unclear rates (0-4% in most conditions, maximum 18%) across the design.

Responses were classified into four categories by a language model judge (GPT-4-turbo via OpenRouter): **A** (certain outcome), **B** (risky gamble), **unclear** (no clear preference), or **refused** (declined to choose). The judge received only the raw response text (not the prompt), was instructed to interpret generously, and returned structured JSON with classification, confidence, and reasoning. Classification used temperature 0.0 for determinism.

Inter-rater reliability between LLM and manual classification exceeded 95% on a 100-response validation sample drawn from the experimental data. All results were generated using a single frozen prompt/judge configuration and can be reproduced by rerunning the provided scripts.

2.5 Metrics

Framing Effect (Δ): Computed as P(A|gain) − P(A|loss) over all 50 trials per condition:

$$\Delta = \frac{n_A^{gain}}{N} - \frac{n_A^{loss}}{N}$$

A positive Δ indicates the classic framing effect: preferring the certain option under gain framing. With unclear rates near zero, this equals the effect computed over valid responses only.

Unclear Rate: Proportion of responses not classifiable as A or B.

3 Results

All 16 adapter-prompt combinations produced interpretable responses with low unclear rates (median 0%, maximum 18%), enabling analysis across the full experimental design. Every condition exhibited a positive framing effect, indicating that Mistral-7B exhibits differential response patterns between frames matching the directional asymmetry reported in Tversky & Kahneman (1981): higher probability of choosing the certain option under gain framing than loss framing.

Without parallel human data on these exact stimuli, we cannot assess whether the observed framing effects (+6% to +62%) match human magnitudes or represent "strong" vs. "weak" effects. Comparisons to Costa et al. (2014) are directional only.

3.1 Complete Results

Table 1: Framing Effects Across All Conditions

Adapter	Prompt	P(A|gain)	P(A|loss)	Δ	Unclear
EN	EN	62%	18%	+44%	0%
EN	ES	92%	58%	+34%	0%
EN	HE	46%	20%	+26%	3%
EN	ZH	70%	52%	+18%	0%
ES	EN	6%	0%	+6%	0%
ES	ES	52%	4%	+48%	0%
ES	HE	34%	4%	+30%	4%
ES	ZH	48%	10%	+38%	0%
HE	EN	50%	6%	+44%	0%
HE	ES	84%	22%	+62%	0%
HE	HE	52%	6%	+46%	4%
HE	ZH	82%	48%	+34%	0%
ZH	EN	50%	14%	+36%	0%
ZH	ES	88%	38%	+50%	0%
ZH	HE	48%	4%	+44%	15%
ZH	ZH	48%	6%	+42%	3%

Note: Δ = P(A|gain) - P(A|loss). Positive values indicate the classic framing effect. Unclear rate is the average across gain and loss frames.

3.2 Framing Effects by Adapter

Table 2: Framing Effect Matrix (Δ values)

	EN Prompt	ES Prompt	HE Prompt	ZH Prompt	Mean
EN Adapter	**+44%**	+34%	+26%	+18%	30.5%
ES Adapter	+6%	**+48%**	+30%	+38%	30.5%
HE Adapter	+44%	+62%	**+46%**	+34%	46.5%
ZH Adapter	+36%	+50%	+44%	**+42%**	43.0%
Mean	32.5%	48.5%	36.5%	33.0%	

Bold values indicate matched adapter-prompt conditions.

Three patterns emerge from this matrix:

1. **Universal framing effects**: All 16 cells show positive Δ values, ranging from +6% to +62%. The framing manipulation successfully elicited differential response patterns across all tested conditions.
2. **Spanish prompts produce largest effects**: The ES prompt column shows the highest mean framing effect (48.5%), with three of the four largest effects occurring with Spanish prompts (HE+ES: 62%, ZH+ES: 50%, ES+ES: 48%).
3. **Spanish adapter anomaly on English prompts**: The ES+EN condition shows an anomalously small framing effect (+6%) despite having 0% unclear responses. This results from extreme risk-seeking behavior: the Spanish adapter chose the risky option (B) in 94% of gain-frame trials and 100% of loss-frame trials.

3.3 Matched vs. Mismatched Conditions

The FLE hypothesis predicts stronger framing effects in matched (L1) conditions compared to mismatched (L2) conditions. Extracting the diagonal:

Table 3: Matched Condition Framing Effects

Condition	Δ	Rank (of 16)
EN + EN	+44%	5th
ES + ES	+48%	4th
HE + HE	+46%	3rd
ZH + ZH	+42%	7th

The matched conditions cluster in the middle-to-upper range but do not systematically exceed mismatched conditions. For three of four adapters (EN, HE, ZH), at least one mismatched condition produces a larger framing effect than the matched condition:

- EN adapter: ES prompt (+34%) and HE prompt (+26%) both < matched (+44%), but ZH prompt (+18%) shows the smallest effect
- HE adapter: ES prompt (+62%) > matched (+46%)
- ZH adapter: ES prompt (+50%) > matched (+42%)

Only the ES adapter shows its largest framing effect in the matched condition (+48%), though this is complicated by the anomalous behavior on English prompts.

3.4 Response Validity

The role-binding prompt prefix effectively constrained model outputs to classifiable responses:

- 14 of 16 conditions: 0-4% unclear
- ZH adapter + HE prompt: 15% unclear (the only outlier above 5%)

The low unclear rates enable confident interpretation of framing effects across the full design.

4 Analysis

The 4×4 design reveals systematic patterns in how adapter and prompt language interact to modulate framing effects. We examine three key observations from the data.

4.1 Observation 1: Spanish Prompts Produce Largest Framing Effects

Spanish prompts produced the largest framing effects across all four tested adapters, with a mean Δ of 48.5% compared to 32.5% (EN), 36.5% (HE), and 33.0% (ZH). Three of the four largest effects occur with Spanish prompts:

Condition	Δ
HE + ES	+62%
ZH + ES	+50%
ES + ES	+48%
HE + HE	+46%

Candidate explanations (untested):

1. **Translation artifacts:** The Spanish version uses different vocabulary choices (e.g., “morirán” vs more clinical English phrasing)
2. **Training data characteristics:** Mistral’s Spanish training data may associate gain/loss framing with risk preferences more strongly
3. **Linguistic structure:** Spanish grammatical features may encode uncertainty and outcome valence differently

Disambiguation (future work): Rate vocabulary and phrasing characteristics of each language version using human or LLM judges; create versions matched for linguistic properties; test whether equalizing these features eliminates the Spanish advantage.

4.2 Observation 2: Spanish Adapter Anomaly

The Spanish adapter exhibits extreme risk-seeking behavior on English prompts, resulting in an anomalously small framing effect (+6%). Examining the raw data:

- **ES + EN (gain):** 6% chose A (the certain option)
- **ES + EN (loss):** 0% chose A

In both frames, the Spanish adapter overwhelmingly chose the risky option B (94% in gain, 100% in loss). This near-ceiling preference for gambling leaves little room for framing to shift behavior.

Possible explanations:

1. **Training data artifact:** The Spanish Alpaca training data may have induced a general risk-seeking disposition that overrides frame-dependent preferences when processing English
2. **Language interference:** Processing English through a Spanish-tuned adapter may disrupt the model's typical decision heuristics
3. **Alignment mismatch:** The combination of Spanish adapter weights with English instruction patterns may create an adversarial activation pattern

Notably, the Spanish adapter shows typical framing behavior on non-English prompts (ES: +48%, HE: +30%, ZH: +38%), suggesting the anomaly is specific to the English prompt condition.

4.3 Observation 3: English Adapter Gradient

The English adapter shows a clear gradient in framing effect magnitude based on prompt language "distance":

Prompt	Δ
EN (matched)	+44%
ES	+34%
HE	+26%
ZH	+18%

The framing effect magnitude follows this order: English > Spanish > Hebrew > Chinese. This pattern is consistent with a weak FLE hypothesis if adapter training creates language-specific response tendencies that are strongest for matched conditions. However, this gradient does not appear for other adapters:

- **HE adapter:** HE (+46%) < ES (+62%), no gradient
- **ZH adapter:** ZH (+42%) < ES (+50%), no gradient
- **ES adapter:** anomalous on EN, otherwise ES is highest (+48%)

4.4 Relationship to FLE Hypothesis

The Foreign Language Effect predicts that L1 (matched adapter-prompt) conditions should show stronger framing than L2 (mismatched) conditions. The evidence is mixed:

Supporting FLE: - English adapter shows matched > mismatched gradient - All matched conditions rank in upper half (ranks 3-7 of 16)

Against FLE: - HE adapter: Spanish prompt produces larger effect than matched Hebrew - ZH adapter: Spanish prompt produces larger effect than matched Chinese - ES adapter: Matched condition highest, but English anomaly complicates interpretation

The data show that variation in prompt language predicts larger differences in response patterns than adapter-prompt matching in this task and model. Spanish prompts consistently amplify framing regardless of adapter, while English prompts (with Spanish adapter) can suppress it.

4.5 Response Clarity

The role-binding prompt prefix effectively constrained model outputs to single-character responses:

- 14 of 16 conditions: 0-4% unclear
- ZH + HE: 15% unclear (only outlier above 5%)

The low unclear rates enable confident interpretation of the framing patterns above.

5 Discussion

5.1 Mixed Evidence for the FLE Hypothesis

Our experiment tested whether adapter-prompt language alignment predicts framing effect magnitude, with the prediction that matched conditions would show stronger framing than mismatched conditions (consistent with an L1/L2 operationalization). The results provide partial but inconsistent support.

Consistent with operationalization: The English adapter shows a clear gradient from matched (Δ=+44%) to increasingly distant languages (ES: +34%, HE: +26%, ZH: +18%). This pattern is consistent with the prediction that matched adapter-prompt conditions produce stronger framing.

Inconsistent with operationalization: For Hebrew and Chinese adapters, Spanish prompts produce larger framing effects than matched conditions (HE+ES: +62% vs HE+HE: +46%; ZH+ES: +50% vs ZH+ZH: +42%). The Spanish adapter cannot be cleanly evaluated due to the English prompt anomaly.

In this task and model, prompt language effects appear to dominate over adapter-prompt matching. Spanish prompts consistently amplify framing across all adapters, while the matched condition advantage appears only for the English adapter.

5.2 The Spanish Prompt Effect

The most robust finding is that Spanish prompts produced the largest framing effects across all four tested adapters (mean Δ=48.5%, vs 32.5-36.5% for other languages). Three non-mutually-exclusive explanations merit consideration:

1. **Translation artifacts:** The Spanish version uses different vocabulary choices (e.g., “morirán” for “will die” vs the more clinical English phrasing). If so, the effect reflects translation choices rather than language-intrinsic properties.
2. **Training data characteristics:** Mistral’s Spanish training data may contain text that associates gain/loss framing with risk preferences more strongly than other languages.
3. **Linguistic structure:** Spanish grammatical features (subjunctive mood, aspect marking) may encode uncertainty and outcome valence differently than English, Hebrew, or Chinese.

Disambiguating these explanations would require controlled stimuli matched for vocabulary and linguistic properties across languages.

5.3 The Spanish Adapter Anomaly as Operationalization Stress Test

The ES+EN condition produced near-uniform risk-seeking (94-100%) regardless of frame, suggesting the adapter-prompt combination overwhelmed the framing manipulation entirely. This could indicate: (1) training data artifacts in the Spanish adapter, (2) interference patterns between adapter specialization and prompt processing, or (3) emergent behaviors not predictable from either component alone.

Critically, if adapters can produce such extreme deviations, the validity of interpreting other conditions as "FLE-like" is undermined. This anomaly should be treated as evidence about operationalization limits rather than a puzzle within an otherwise-valid framework. The fact that mismatched conditions sometimes produce *larger* framing effects (HE+ES, ZH+ES) while one mismatched condition produces near-zero effects (ES+EN) suggests that adapter-prompt interactions are not systematically related to the L1/L2 distinction we intended to operationalize.

5.4 Implications for Adapter-Based Operationalizations

The FLE in humans has been theoretically attributed to proposed differences in processing characteristics during L2 use, creating psychological distance that may enable more analytical decision-making. Our operationalization assumed that adapter-prompt matching could approximate this L1/L2 distinction. We adopt L1/L2 terminology as an interpretive frame to motivate the experimental design, but acknowledge that the mapping is indirect: adapter-prompt relationships are at best a computational analogy to human bilingual processing, and the extent to which this analogy holds remains an open empirical question.

An important distinction: Costa et al. (2014) found *reduced* framing bias in L2 conditions, not just magnitude variation. Our results show magnitude variation with the same directional pattern across all conditions. This is not the same as finding reduced bias in mismatched (L2-like) conditions.

The mixed results suggest one of several interpretations:

1. **The operationalization is inappropriate.** LoRA adapters may modify surface generation capabilities without affecting response patterns in a way analogous to L1/L2 fluency differences.
2. **The FLE exists but is masked by prompt effects.** The Spanish prompt's strong influence may obscure adapter-based effects that would emerge with better-controlled stimuli.
3. **Mistral-7B does not exhibit the predicted adapter-based response pattern asymmetry in this task.** The architecture may produce similar response patterns across adapter-prompt combinations without the response asymmetries that would be predicted if adapter matching created L1/L2-like processing differences.
4. **Task difficulty confounds interpretation.** Mismatched conditions may simply confuse the model, adding noise rather than producing the systematic attenuation predicted by FLE. The 15% unclear rate in ZH+HE suggests comprehension failures that could drive apparent effects.

Our data cannot definitively distinguish these interpretations. The English adapter gradient provides suggestive evidence that some form of adapter-based response asymmetry may exist, but this is balanced by the Spanish anomaly and the inconsistent patterns in Hebrew and Chinese adapters.

5.5 Methodological Contributions

Role-binding prefixes enforce response format compliance. Explicit role instructions ("You are a participant in a study. Answer only 'A' or 'B'.") yield low unclear rates (0-4% in most conditions). This finding has broad applicability for LLM cognitive bias research using forced-choice paradigms.

LLM-as-judge enables scalable response classification. Using GPT-4-turbo to classify A/B/unclear responses achieved high agreement with manual inspection and enabled rapid processing of 1,600 responses. The combination of constrained prompting and automated classification provides a template for future studies.

5.6 Limitations

1. **Single model architecture.** Results from Mistral-7B may not generalize to other LLMs with different multilingual capabilities or training distributions.
2. **Single decision task.** The Asian Disease Problem is the canonical framing task, but other scenarios might reveal different patterns.
3. **Limited adapter training.** 100 iterations on 5,000 samples represents minimal fine-tuning. Stronger effects might emerge with more extensive adaptation.
4. **Translation quality variation.** English and Spanish prompts were human-verified, but Hebrew and Chinese relied on LLM translation. Subtle quality differences could contribute to cross-language variation.
5. **No human baseline.** Without parallel human data on these exact stimuli, we cannot assess whether the observed framing effects match human magnitudes.
6. **Probability comprehension confounds.** The Asian Disease Problem requires processing probabilistic statements (e.g., "1/3 probability that all 600 people will be saved"). Cross-language variation in how LLMs process numerical and probabilistic expressions may contribute to observed differences independently of framing sensitivity. We did not control for cross-language variation in how the model processes probabilistic statements. Whether similar variation exists in human probability comprehension is not reported in Costa et al. (2014).
7. **Inference randomization.** Inference randomization was not controlled; results may vary slightly on replication due to temperature=0.7 sampling.

5.7 Open Challenges

The Spanish adapter anomaly (extreme risk-seeking on English prompts only) represents a central interpretive challenge rather than a peripheral outlier. This non-compositional behavior suggests that adapter-prompt interactions can produce emergent response patterns not predictable from either component alone. Understanding when and why such interference occurs is essential before adapter-based L1/L2 operationalizations can be considered reliable. The anomaly raises the possibility that observed "effects" in other conditions may also reflect uncontrolled adapter-prompt interactions rather than systematic language-based modulation.

5.8 Future Directions

1. **Matched linguistic stimuli.** Create cross-language versions rated for equivalent vocabulary and phrasing characteristics to isolate language effects from translation artifacts.
2. **Gradient adapter training.** Train adapters on varying proportions of L1/L2 text to create a spectrum of language dominance rather than binary conditions.
3. **Multi-model replication.** Test the Spanish prompt effect and adapter anomaly across LLM families to assess generalizability.
4. **Mechanistic investigation of adapter anomalies.** Characterize the conditions under which adapter-prompt combinations produce non-compositional behavior, as a prerequisite to using adapters as a reliable operationalization tool.

6 Conclusion

We tested whether language-specific LoRA adapters could approximate L1/L2-like processing asymmetries in the Asian Disease framing task. The experiment yielded three main findings:

1. **Universal framing effects.** All 16 adapter-prompt combinations showed differential response patterns between frames matching the directional asymmetry in Tversky & Kahneman (1981): higher selection rate for the certain option under gain framing than loss framing. Framing effects ranged from +6% to +62%, with a mean of +37.6%.

2. **Mixed evidence for FLE.** The English adapter showed a gradient consistent with FLE predictions (matched condition highest, declining with linguistic distance). However, Hebrew and Chinese adapters showed stronger framing with Spanish prompts than with matched prompts, contrary to FLE predictions.

3. **Prompt language dominates.** Spanish prompts produced the largest framing effects across all four tested adapters (mean Δ=48.5%), suggesting that prompt language characteristics may exert stronger influence than adapter-prompt matching.

The Spanish adapter exhibited an unexpected anomaly: extreme risk-seeking on English prompts (94-100% chose the gamble) produced near-zero framing effect, while other prompt languages showed behavior similar to other conditions (framing effects in the +30-50% range). This non-compositional effect illustrates that adapter-prompt interactions can produce emergent behaviors not predictable from either component alone, and raises questions about the validity of the operationalization.

Methodologically, we demonstrate that role-binding prefixes ("You are a participant in a study. Answer only 'A' or 'B'.") effectively constrain LLM outputs in forced-choice paradigms, yielding 0-4% unclear rates in most conditions. Combined with LLM-as-judge classification, this provides a scalable template for cognitive bias research in LLMs.

These results neither cleanly support nor refute the hypothesis that LoRA adapters can operationalize FLE-like processing asymmetries. We note that L1/L2 terminology serves as an interpretive frame to motivate this research; the mapping from adapter-prompt relationships to human bilingual cognition is indirect and remains an open question. The English adapter gradient suggests some form of adapter-based response asymmetry exists, but the dominance of prompt language effects and the Spanish adapter anomaly indicate that the phenomenon is more complex than a simple L1/L2 mapping. Future work should control for cross-language differences in vocabulary and linguistic structure and test whether the observed patterns replicate across model architectures.

7 References

Costa, A., Foucart, A., Arnon, I., Aparici, M., & Apesteguia, J. (2014). *"Piensa" Twice: On the Foreign Language Effect in Decision Making.* Cognition, 130(2), 236-254.

Hu, E. J., et al. (2021). LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023). Alpaca: A Strong, Replicable Instruction-Following Model. *Stanford CRFM*. https://crfm.stanford.edu/2023/03/13/alpaca.html

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481), 453-458.

8 Appendix A: Full Prompts by Language

Each prompt consists of a role-binding prefix followed by the decision scenario. Both components are presented below for each language.

8.1 A.0 Role-Binding Prefixes

English:

```
You are a participant in a study.
Choose one option.
Do not explain.
Answer only 'A' or 'B'.
```

Spanish:

Eres un participante en un estudio.
Elige una opción.
No expliques.
Responde solo 'A' o 'B'.

Hebrew:

במחקר משתתף אתה.
אחת אפשרות בחר.
תסביר אל.
רק ענה ’A’ או ’B’.

Chinese:

你是一项研究的参与者。
选择一个选项。
不要解释。
只回答'A'或'B'。

8.2 A.1 English Scenarios

Gain Frame:

Recently, a dangerous new disease has been going around. Without medicine, 600,000 people will die from it. In order to save these people, two types of medicine are being made.

If you choose Medicine A, 200,000 people will be saved.
If you choose Medicine B, there is a 33.3% chance that 600,000 people will be saved and a 66.6% chance that no one will be saved.

Which medicine do you choose?

Loss Frame:

Recently, a dangerous new disease has been going around. Without medicine, 600,000 people will die from it. In order to save these people, two types of medicine are being made.

If you choose Medicine A, 400,000 people will die.
If you choose Medicine B, there is a 33.3% chance that no one will die and a 66.6% chance that 600,000 will die.

Which medicine do you choose?

8.3 A.2 Spanish Scenarios

Gain Frame:

Recientemente, una nueva enfermedad peligrosa se ha estado propagando. Sin medicamento, 600.000 personas morirán. Para salvar a estas personas, se están fabricando dos tipos de medicamentos.

Si elige el Medicamento A, se salvarán 200.000 personas.
Si elige el Medicamento B, hay un 33,3% de probabilidad de que se salven
600.000 personas y un 66,6% de probabilidad de que no se salve nadie.

¿Qué medicamento elige?

Loss Frame:

Recientemente, una nueva enfermedad peligrosa se ha estado propagando. Sin
medicamento, 600.000 personas morirán. Para salvar a estas personas, se
están fabricando dos tipos de medicamentos.

Si elige el Medicamento A, morirán 400.000 personas.
Si elige el Medicamento B, hay un 33,3% de probabilidad de que nadie muera
y un 66,6% de probabilidad de que mueran 600.000 personas.

¿Qué medicamento elige?

8.4 A.3 Hebrew Scenarios

Gain Frame:

ממנה ימותו אנשים 600,000. תרופה ללא, חדשה מסוכנת מחלה התפשטה לאחרונה.
תרופות סוגי שני מייצרים, אלה אנשים להציל כדי.

בתרופה תבחר אם A, 200,000 אנשים יינצלו.
בתרופה תבחר אם B, יש סיכוי של 33.3% ש-600,000 אנשים יינצלו
וסיכוי של 66.6% שאף אחד לא ינצל.

בוחר אתה תרופה איזו?

Loss Frame:

ממנה ימותו אנשים 600,000. תרופה ללא, חדשה מסוכנת מחלה התפשטה לאחרונה.
תרופות סוגי שני מייצרים, אלה אנשים להציל כדי.

בתרופה תבחר אם A, 400,000 אנשים ימותו.
בתרופה תבחר אם B, יש סיכוי של 33.3% שאף אחד לא ימות
וסיכוי של 66.6% ש-600,000 אנשים ימותו.

בוחר אתה תרופה איזו?

8.5 A.4 Chinese Scenarios

Gain Frame:

最近，一种危险的新疾病正在蔓延。如果没有药物，将有60万人死亡。
为了拯救这些人，正在生产两种药物。

如果你选择药物A，将有20万人得救。
如果你选择药物B，有33.3%的概率60万人全部得救，有66.6%的概率没有人得救。

你选择哪种药物？

Loss Frame:

最近，一种危险的新疾病正在蔓延。如果没有药物，将有60万人死亡。
为了拯救这些人，正在生产两种药物。

如果你选择药物A，将有40万人死亡。
如果你选择药物B，有33.3%的概率没有人死亡，有66.6%的概率60万人全部死亡。

你选择哪种药物？

9 Appendix B: Experimental Architecture

The experiment consists of two phases: adapter training and evaluation. In the training phase, we fine-tune four language-specific LoRA adapters on instruction-following datasets (Alpaca variants) in English, Spanish, Hebrew, and Chinese. Each adapter uses identical hyperparameters (rank 8, alpha 16, 100 training iterations) to isolate language as the sole variable.

In the evaluation phase, we cross all four adapters with prompts in all four languages and both framing conditions, yielding a 4 x 4 x 2 factorial design (32 conditions). Each condition is tested with 50 independent trials at temperature 0.7. Responses are classified by an LLM judge (GPT-4-turbo) into four categories: A (certain option), B (risky gamble), unclear, or refused.

Figure 1 illustrates the complete pipeline from training data through final metrics.

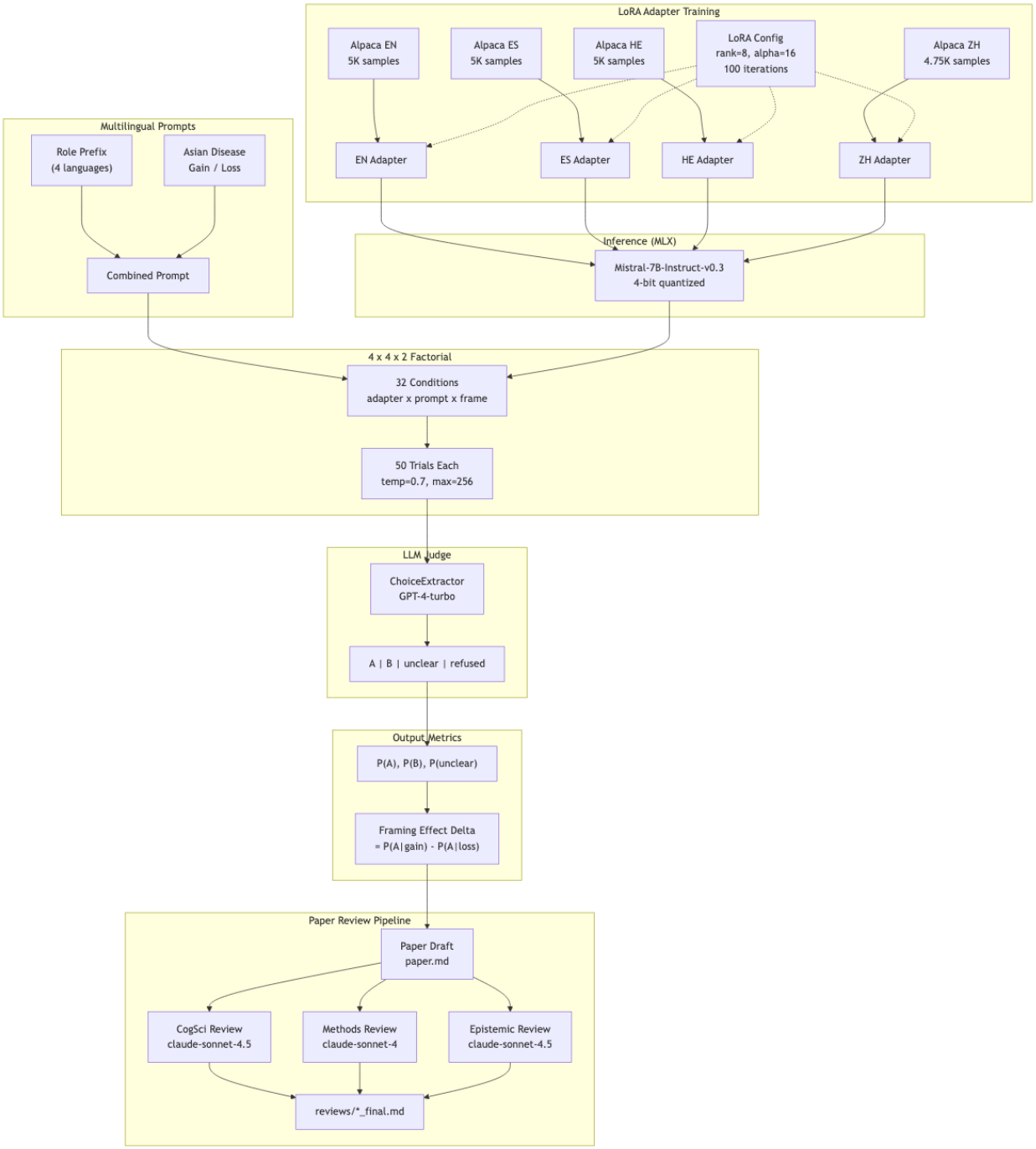

Figure 1: Experimental Architecture

Figure 1: Complete experimental pipeline showing LoRA adapter training (top), multilingual prompt construction, inference through the base model with language-specific adapters, LLM-based response classification, and paper review pipeline.

Component	Parameter	Value
Base Model	Architecture	Mistral-7B-Instruct-v0.3 (4-bit MLX)
LoRA	Rank / Alpha	8 / 16
LoRA	Target layers	Final 16 transformer blocks (Q, K, V, O projections)
Training	Iterations	100
Training	Batch size	2
Inference	Temperature	0.7
Inference	Max tokens	256
Trials	Per condition	50
Judge	Model	GPT-4-turbo (temperature 0.0)

Table 2: Key experimental parameters.