Main Hypotheses & Results

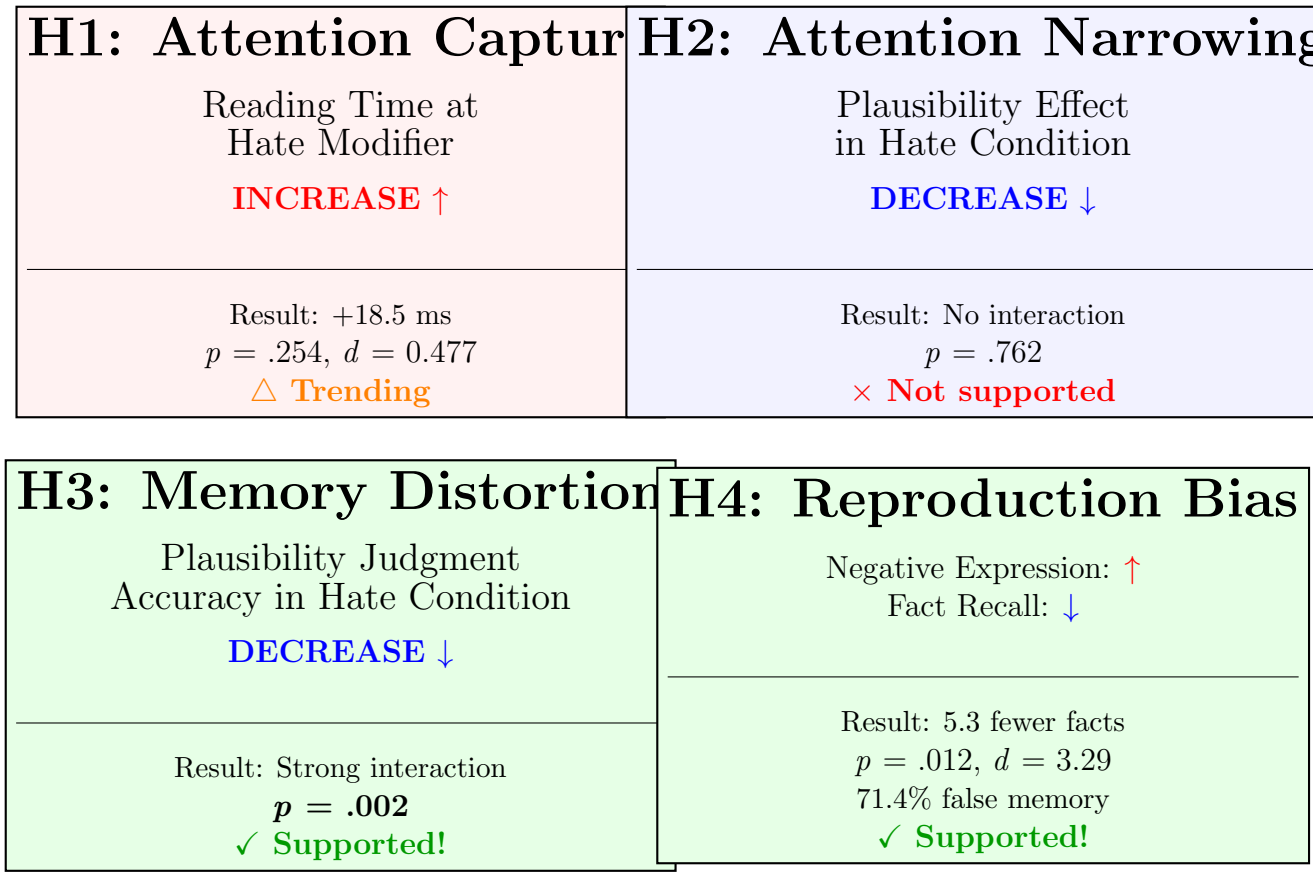

Figure 1: Overview of four main hypotheses with predicted effects and empirical results. Red arrows (↑) indicate predicted increases; blue arrows (↓) indicate predicted decreases.

Detailed Hypothesis Descriptions

H1: Attention Capture

Prediction: Hate modifiers will elicit longer reading times (↑) than neutral modifiers, reflecting affect-driven attentional capture.

Result:

- Original data: +7.2 ms, $p = .468$, $d = 0.293$
- With outlier removal (200–1600ms): +18.5 ms, $p = .254$, $d = 0.477$

Status: Trending effect in predicted direction (effect size increases 63% with stricter outlier criteria)

Interpretation: Direction consistent with hypothesis. Single outlier (1725 ms) substantially influenced results, demonstrating importance of data quality control.

H2: Attention Narrowing & Shallow Integration

Prediction:

- Neutral-modifier sentences: Clear plausibility effect (Implausible ¿ Plausible RT)
- Hate-modifier sentences: Reduced plausibility effect (↓), indicating shallower semantic integration under attentional narrowing

Result:

- Neutral context: NI – NP = +7.06 ms
- Hate context: HI – HP = +7.10 ms
- Interaction: $F(1,6) = 0.00$, $p = .995$

Status: Not supported

Interpretation: No evidence of attention narrowing effect. Possible reasons: (1) small sample size (N=7), (2) weak plausibility manipulation, (3) need to examine spillover region.

H3: Biased Memory (Trade-off + Distortion)

Prediction: Relative to neutral context, hate context will lead to lower accuracy (↓) for plausibility discrimination in recognition memory.

Result:

- **Strong Emotion × Plausibility interaction: $p = .002$**
- Neutral condition: Clear discrimination (P – I = +0.593, $p = .001$) ✓
- Hate condition: No discrimination (P – I = –0.171, $p = .439$) ×

Status: **Strongly supported!**

Interpretation: Hate speech disrupts accurate encoding and retrieval of plausibility information. Exact replication of previous dataset ($p = .002$ in both datasets). Distortion index shows 5/7 participants exhibited expected pattern.

H4: Encoding Bias in Reproduction

Prediction: Free descriptions after hate context will contain:

- Higher proportion (↑) of hate-consistent propositions and negative adjectives
- Fewer neutral background details (↓)

Result:

- **Negative expression users recalled 5.3 fewer facts** (2.5 vs. 7.8 facts)
- Independent samples t-test: $t(5) = 3.22$, $\boldsymbol{p = .012}$, Cohen's $d = 3.29$
- **71.4% of participants included false information** (implausible content recalled as fact)
- Mean false information: 2.29 instances per participant

Status: Supported!
Critical Methodological Finding:

- All negative expressions (100%) were *indirect*: "unsophisticated" (*cheonbak*), "ignorant" (*muji*), "low-level" (*sujun nat*)
- Zero direct hate speech reproduced
- Suggests hate speech induces **schema-level implicit bias** rather than explicit word copying
- Social desirability prevents direct hate reproduction, but fundamental negative attitude persists through indirect language

Summary Table

Hypothesis	Measure	Result	p-value	Status
Manipulation	Negativity rating	$d = 4.18$	¡ .0001	✓✓✓
H1 (original)	Modifier RT	+7.2 ms	.468	△
H1 (strict)	Modifier RT	+18.5 ms	.254	△
H2	Interaction	+7.1 ms	.762	×
H3	Interaction	+0.734	**.002**	✓✓
H4	Fact recall diff	–5.3 facts	**.012**	✓✓
H4	False memory	71.4%	—	✓

Table 1: Summary of hypothesis testing results. ✓ = supported, △ = trending, × = not supported