情绪LLM白盒攻击研究

White-box Targeted Emotion Attack & Modality Conflict Mechanism Analysis

徐振宇、李享

目录

1.1 问题定义与核心目标

研究问题

是否可以通过对输入语音进行微小扰动，使多模态语音-语言大模型在理解该语音时，将说话人的情绪稳定判断为 目标情绪 （如 happy），同时保持语义内容基本不变?

攻击约束

- 不改模型参数
- 不改 prompt
- **只优化输入音频**

输出形式

情绪判断任务下输出单词情绪标签：

happy / sad / angry / neutral

1.2 模型抽象与基本表征

条件生成模型（数学抽象表示）

$$p\theta(y \mid x, p)$$

变量含义

x：输入语音信号

p：prompt（离散且固定）

y：生成 token 序列

θ：模型参数（固定，不参与优化）

内部两步

1. 音频编码（共享）

$$h = f_{audio}(x)$$

2. 条件生成（依赖 prompt）

$$p\theta(y \mid x, p) = p\theta(y \mid h, p)$$

重要事实：对同一段音频 x，不同 prompt 下使用同一个音频表示h

1.3 情绪的内部表征

核心观点

- 模型内部**不存在显式"情绪变量"**
- 情绪体现在：在特定任务条件下，对情绪相关 token 的概率分布偏好

情绪判断的分布形式

当 prompt 要求模型输出情绪标签时，在某个生成位置形成分布：

$$p\theta(y1 = v \mid h, pemo)$$
$$v \in \{happy, sad, angry, neutral\}$$

1.4 攻击目标：能量最小化

目标情绪损失（以 happy 为例）

$$\mathcal{L}_{emo}(x') = -\log p_\theta(y_1 = happy \mid h(x'), p_{emo})$$

直观含义

- 若模型**不倾向**输出 happy → 损失大
- 若模型**高度确信** happy → 损失趋近 0

"Happy 区域"概念

$$\mathcal{H}happy = \{h \mid p\theta(happy|h, pemo) > p\theta(v|h, pemo), \forall v \neq happy\}$$

攻击目标：通过优化扰动，使 h(x') 推入 $\mathcal{H}$happy

1.5 语义一致性约束

> **问题**
> 仅优化情绪损失可能出现退化

基准转写定义

$$yasr(x) = argmaxy\, p\theta(y \mid h(x), pasr)$$

语义一致性约束

$$\mathcal{L}asr(x') = -\log p\theta(yasr(x) \mid h(x'), pasr)$$

作用

要求在转写任务上，对抗音频与原始音频具有**同一高概率解释**，从而强制模型继续"认真听音频"，避免破坏语义

1.6 实验结果（初步）

攻击闭环已跑通

输入音频 → 优化扰动 → 输出指定情绪

情绪攻击成功率

80%

prompt：直接输出音频情绪标签（中性）

评估限制

1. WER 等参数设置过于严格，后续批量实验准备接入商业 API 判断
2. 成功率并不精确，需要后续改进方法论后批量实验完善

攻击效果对 prompt 敏感

- 直接分类提示：~40%
- 描述+分类提示：~50%
- 逐步分析提示：~20%

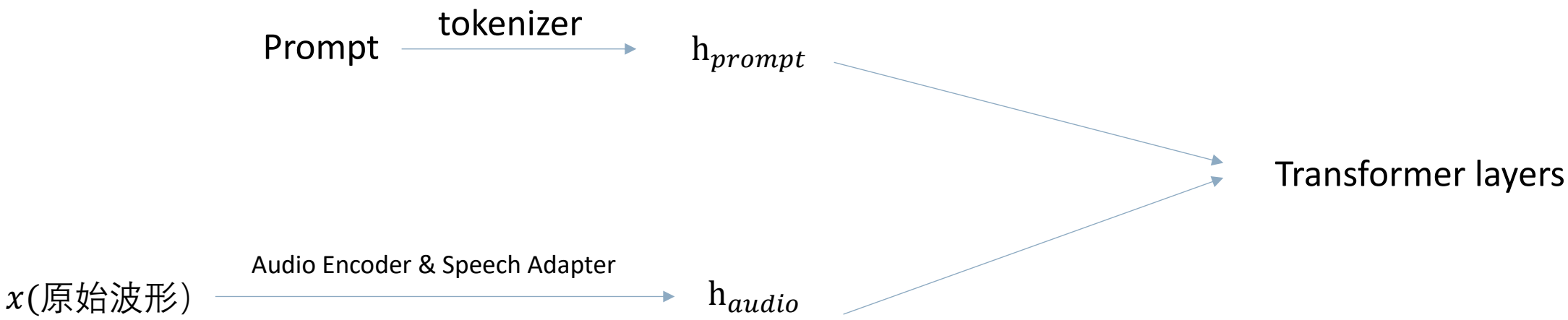
结论

定向情绪攻击可行，但跨 prompt 鲁棒性需要机理分析支撑（缺少why）

02 Audio内部冲突机理

Semantic vs Prosody

2.0 输入格式解析


Prompt
tokenizer
h_{prompt}
Transformer layers
x(原始波形)
Audio Encoder & Speech Adapter
h_{audio}

2.1 动机与核心问题

音频内部两种情绪线索

语义情绪（semantic）

文本内容本身表达的情绪

韵律情绪（prosody）

语调 / 节奏 / 强度等副语言线索

核心问题

1. 在模型不同层里，哪类信息**更强、更可读**?
2. 尤其在语义与韵律**冲突**时，表征更偏向哪一边?

研究目标：通过逐层分析揭示模型内部的情绪信息处理机制

2.2 实验设计：逐层表示 + Probe 可读性

情绪类别（5类）

neutral / happy / sad / angry / surprised

数据

- 50条文本（每类10条）
- 每条生成5种韵律版本
- 理论250条，实际可用247条
- 冲突样本 197；一致样本 50

Prompt 固定为中性判断

"What is the emotion of this audio? Answer with exactly one word: neutral, happy, sad, angry, surprised."

表示提取方式

- 一次 forward 获取 **0–35 层** hidden states
- 取 audio span 做 pooling 得到向量
- 每层训练两个 probe：
  - 预测 text_emotion（语义）
  - 预测 prosody_emotion（韵律）

主导性指标

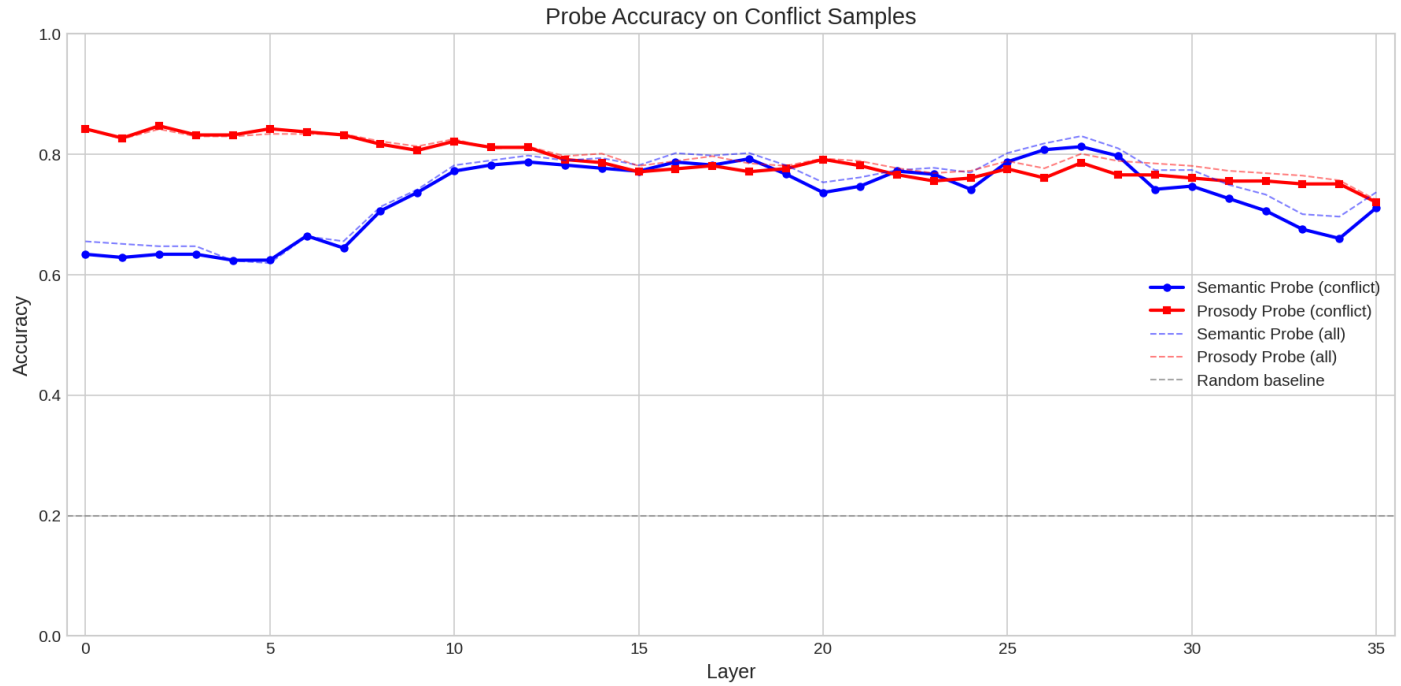
$$D(\ell) = \mathrm{Acc}_{\mathrm{prosody}}(\ell) - \mathrm{Acc}_{\mathrm{semantic}}(\ell)$$

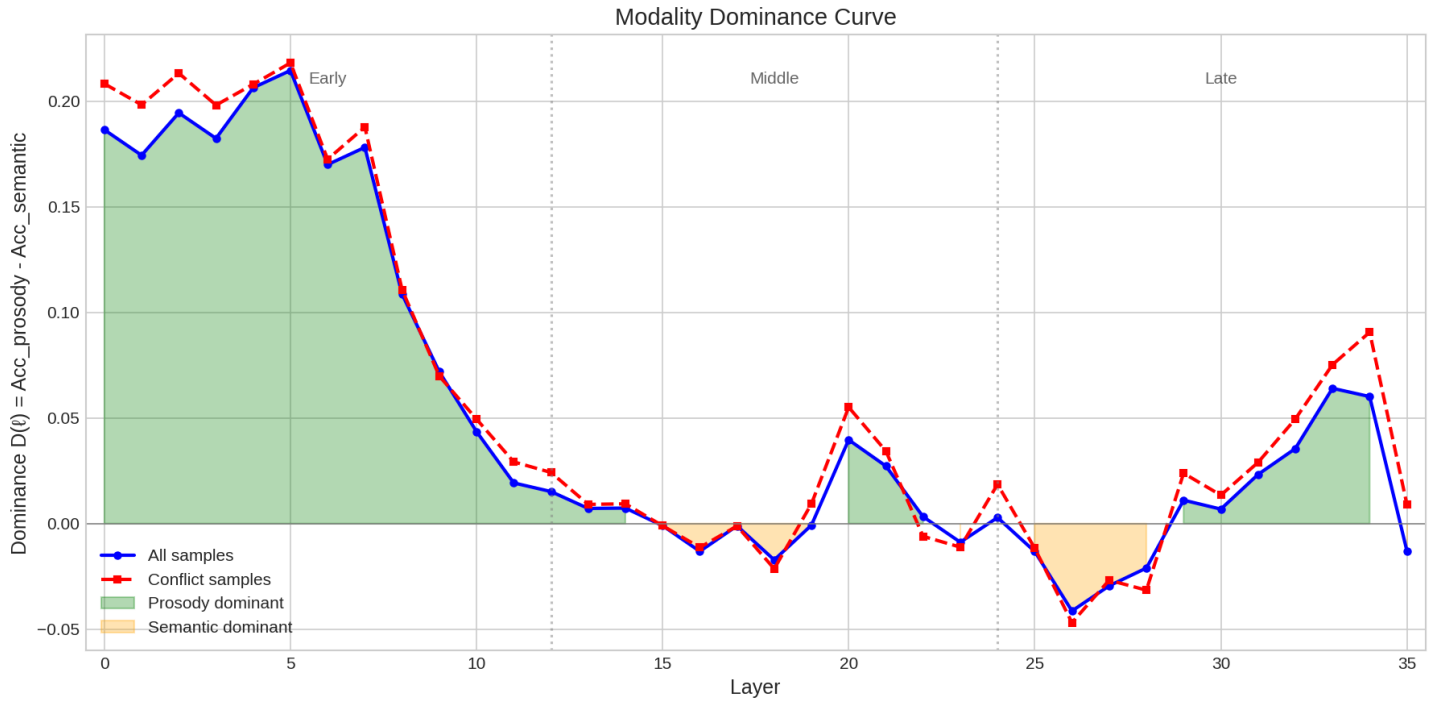
- $D > 0$：韵律更"可读"
- $D < 0$：语义更"可读"

2.3 Probe结果：层级结构

关键数值

overall dominance	prosody
平均 D	≈ 0.0526
韵律最强（layer 0）	prosody_acc ≈ 0.842
语义最强（layer 27）	semantic_acc ≈ 0.830
韵律主导峰值（layer 5）	D ≈ 0.2146
语义占优最明显（layer 26）	D ≈ −0.0414





2.4 冲突子集统计与 Logit Lens

冲突子集（197条）

avg semantic_acc	≈ 0.7299
avg prosody_acc	≈ 0.7895
avg dominance	≈ 0.0596

Logit Lens 方法细节

- 只取冲突样本（semantic_label ≠ prosody_label）
- 一次前向拿到所有层表示：hidden_states[l][t]
- Readout position：取最后一个输入 token（readout_pos = T − 1）
- 使用真实输出路径读出 logits：$\text{logits}_\ell = \text{LMHead}(\text{FinalNorm}(h_\ell))$
- Restricted 5-way：只看 neutral/happy/sad/angry/surprised 的 logits

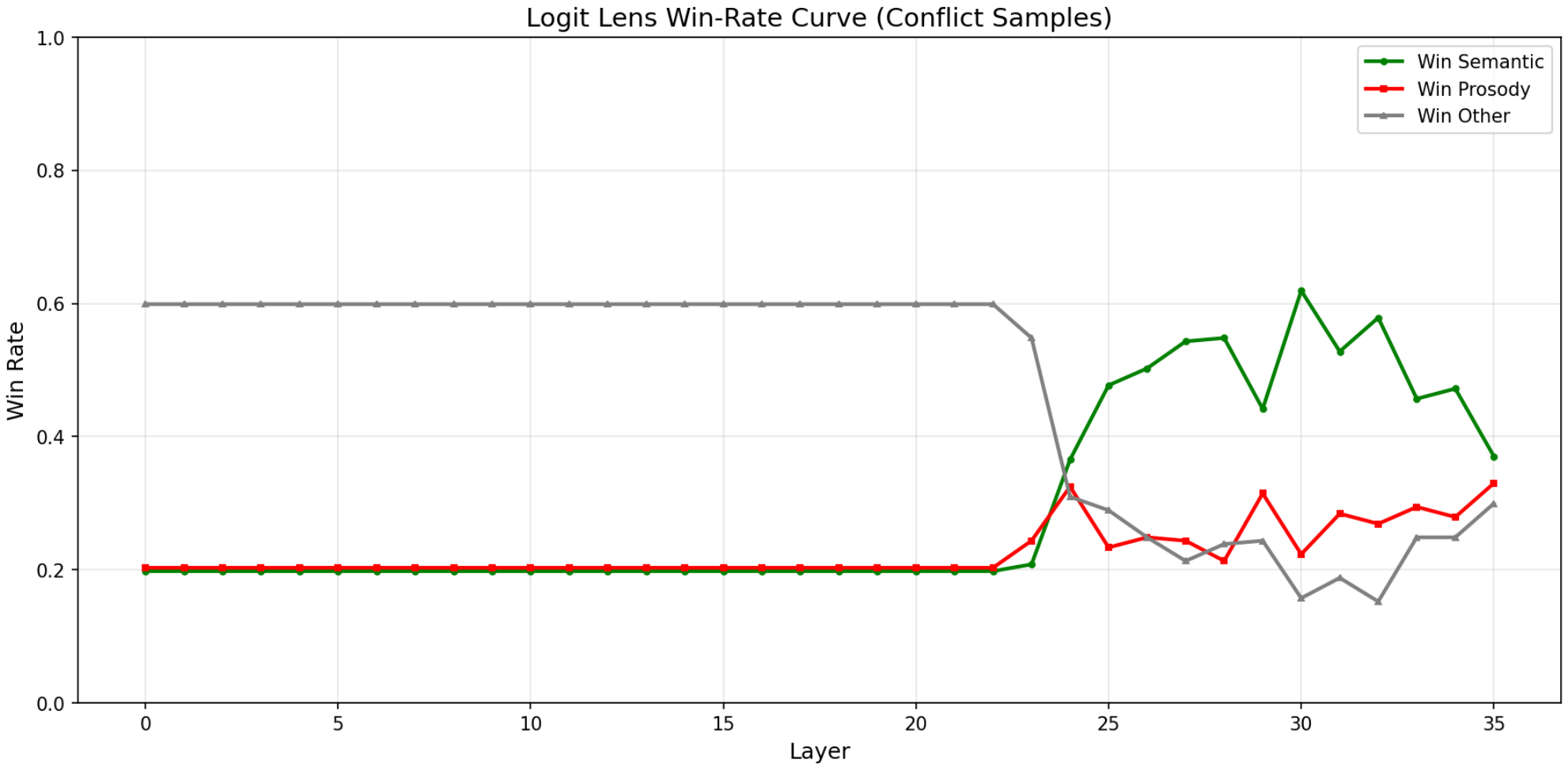
解读

- 层内表征层面：**韵律线索整体更可读**
- 但最终输出是否采用韵律，需要进一步看 决策轨迹/因果证据

主要观察

- 前半段层（0–22）：margin 接近 0（语义 vs 韵律**混合**）
- 后半段层（约23–35）：margin 明显下降为负（逐层更**偏语义**）
- Win-rate：前半段"其他"占比高（不稳定），后半段语义赢率明显上升

2.4 冲突子集统计与 Logit Lens



2.5 Activation Patching：因果证据

基本原理

对样本对 (A,B)，在某层 ℓ 把 A 的 audio 区域表示替换成 B 的，再继续前向到输出。若替换后输出显著朝 B 的目标标签变化，说明该层 audio 表示对该标签具有因果控制力。

指标

- **Flip to Target**：pred_patch(ℓ) == target 的比例
- **Flip from Base**：pred_patch(ℓ) != pred_base 的比例
- **Delta Logit(Target)**：$\Delta(\ell) = \text{logit}_{\text{patch}}(\text{target}) - \text{logit}_{\text{base}}(\text{target})$

Semantic patch（只换语义）

- 早期层（0–12）：flip_to_target **很高**（最高约 0.65，多层维持 0.5+）
- delta_logit(target) 早期**很大**（约 5–6），随后逐步衰减

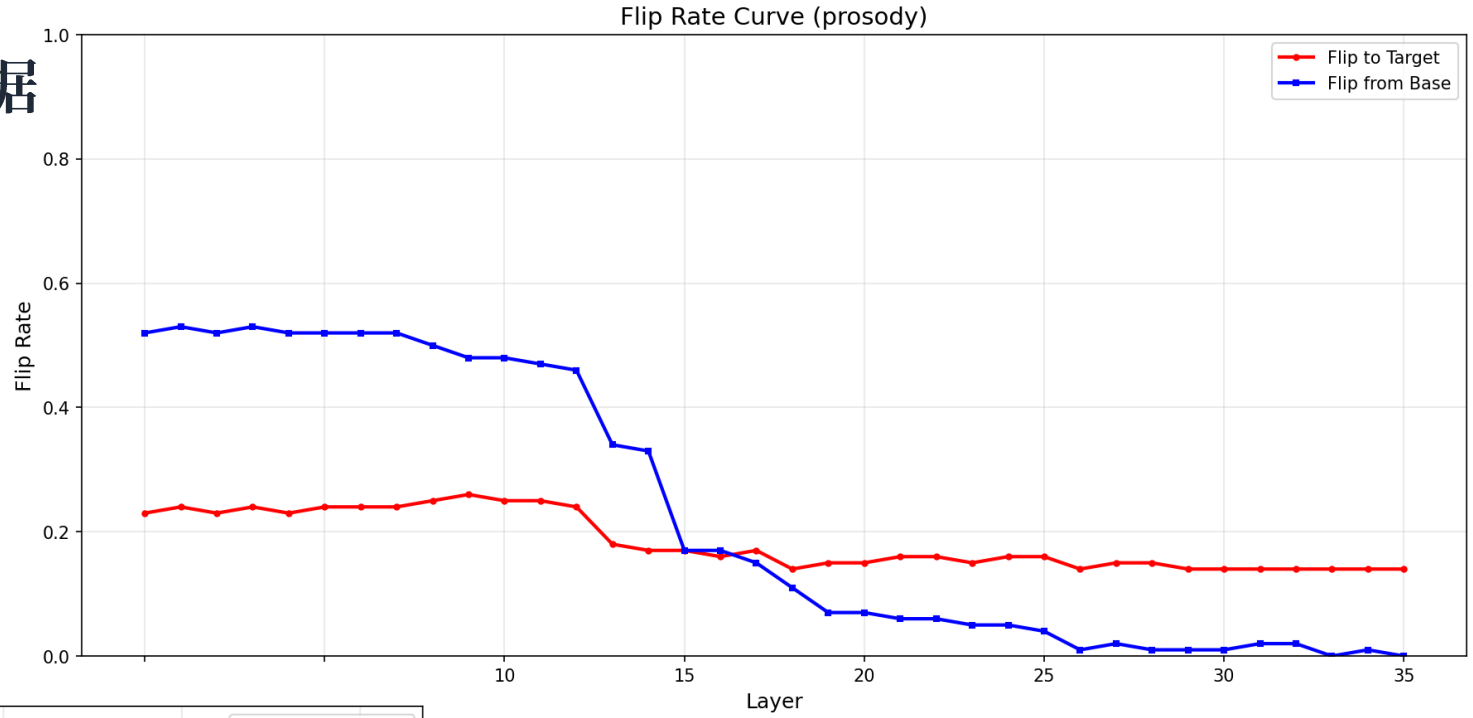
Prosody patch（只换韵律）

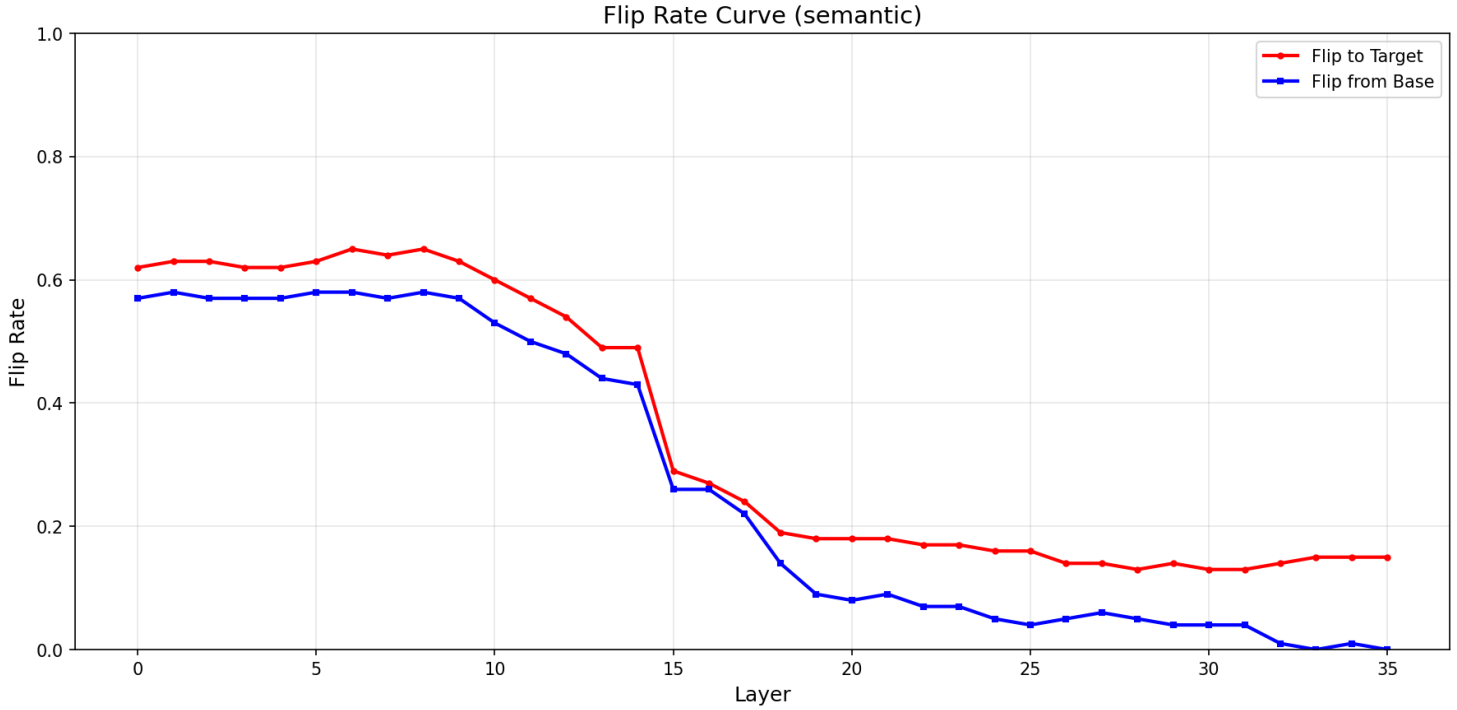
- flip_to_target **整体较低**（约 0.14–0.26），峰值在较早层（~9）
- delta_logit(target) 早期**较小**（~2.2–2.3），随后逐步衰减

关键边界

约 14–15 层是"audio span 可控性"的边界。从这里开始，仅替换 audio span **难以影响最终输出**（信息已扩散到全局/其他位置）

2.5 Activation Patching：因果证据





03 Prompt-Audio冲突机理

Instruction vs Audio

3.1 动机与实验设置

动机（两问）

1 **问题1**
情感信息的消解发生在模型**何处**？（where）

2 **问题2**
谁具有**主导性**因果关系？（who）

数据与控制

- 数据集：2组中英文，5种情绪的对照音频（TTS生成）
- 固定语义文本（避免语义干扰）：

"The package is scheduled to arrive..."

实验分组

组别	音频	文本指令
Audio-only	A	请判断音频情绪
Conflict	A	请以T判断（T ≠ A）
Consistent	A	请以T判断（T = A）

方法（两类）

1) Logit Lens 差分定位 where

逐层比较 T 与 A 的 logit 差值

$$\Delta\text{logit}_\ell = \text{logit}_\ell(\text{T}) - \text{logit}_\ell(\text{A})$$

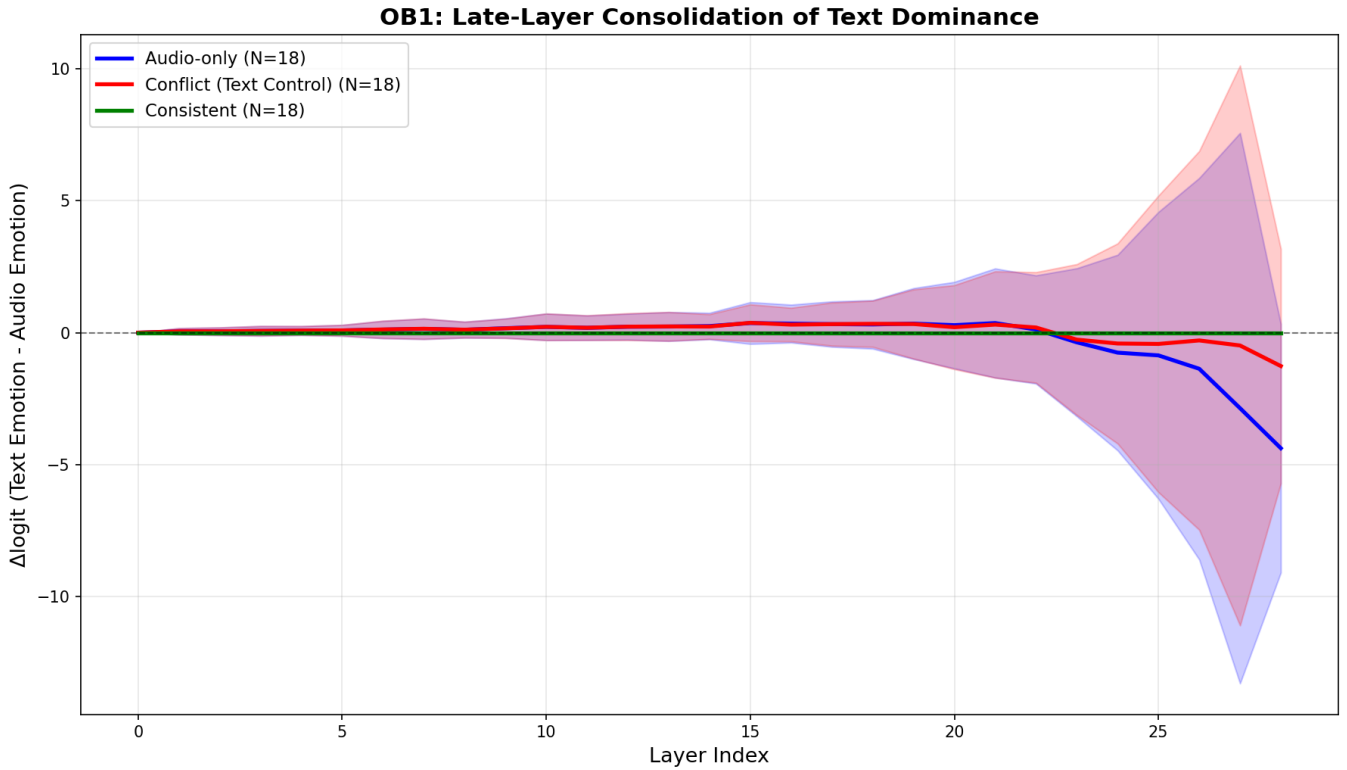
2) Activation patching 定位 who

不同层替换文本/音频 hidden states，对比 PatchText vs PatchAudio

3.2 Finding 1：仲裁发生在晚期层（26–28）

观察结果

通过对三组（Audio-only / Conflict / Consistent）做 Logit Lens 差分观察，临界层基本发生在 26–28 层



3.3 Finding 2：文本因果主导，音频中层被wash-out

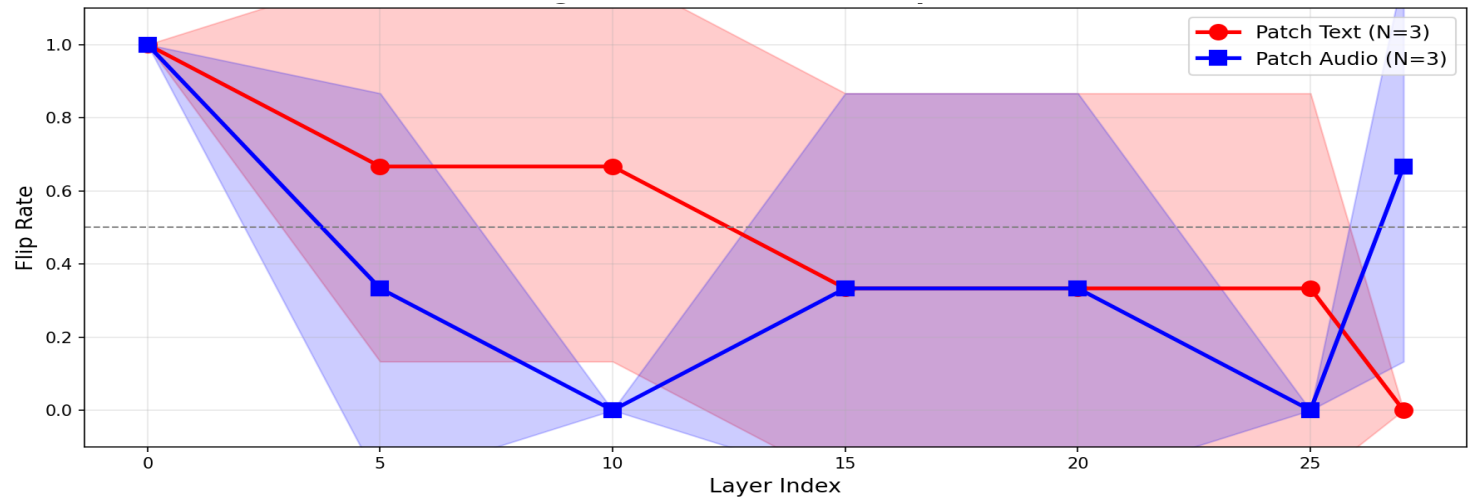
结论

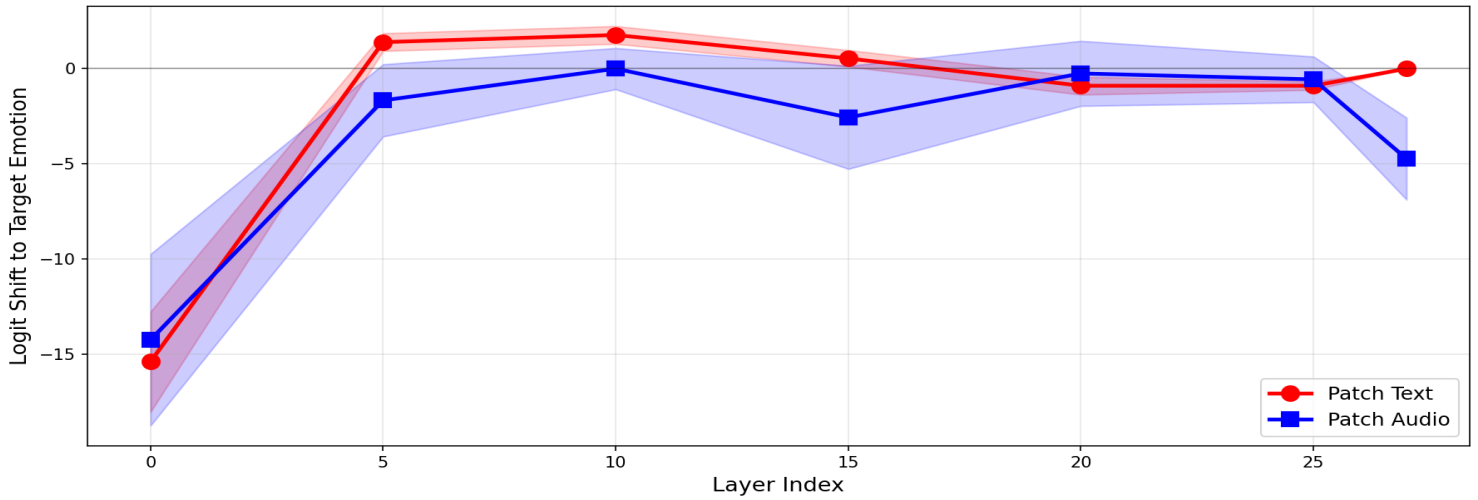
Patch Text 在中层有效

文本指令在决策窗口（5–20）保持 强因果效应

Patch Audio 无效/不稳定

音频信号在中层被 structural wash-out





04 机理引导攻击优化

Mechanism-Guided Attack Optimization

4.1 机理引导攻击优化方向