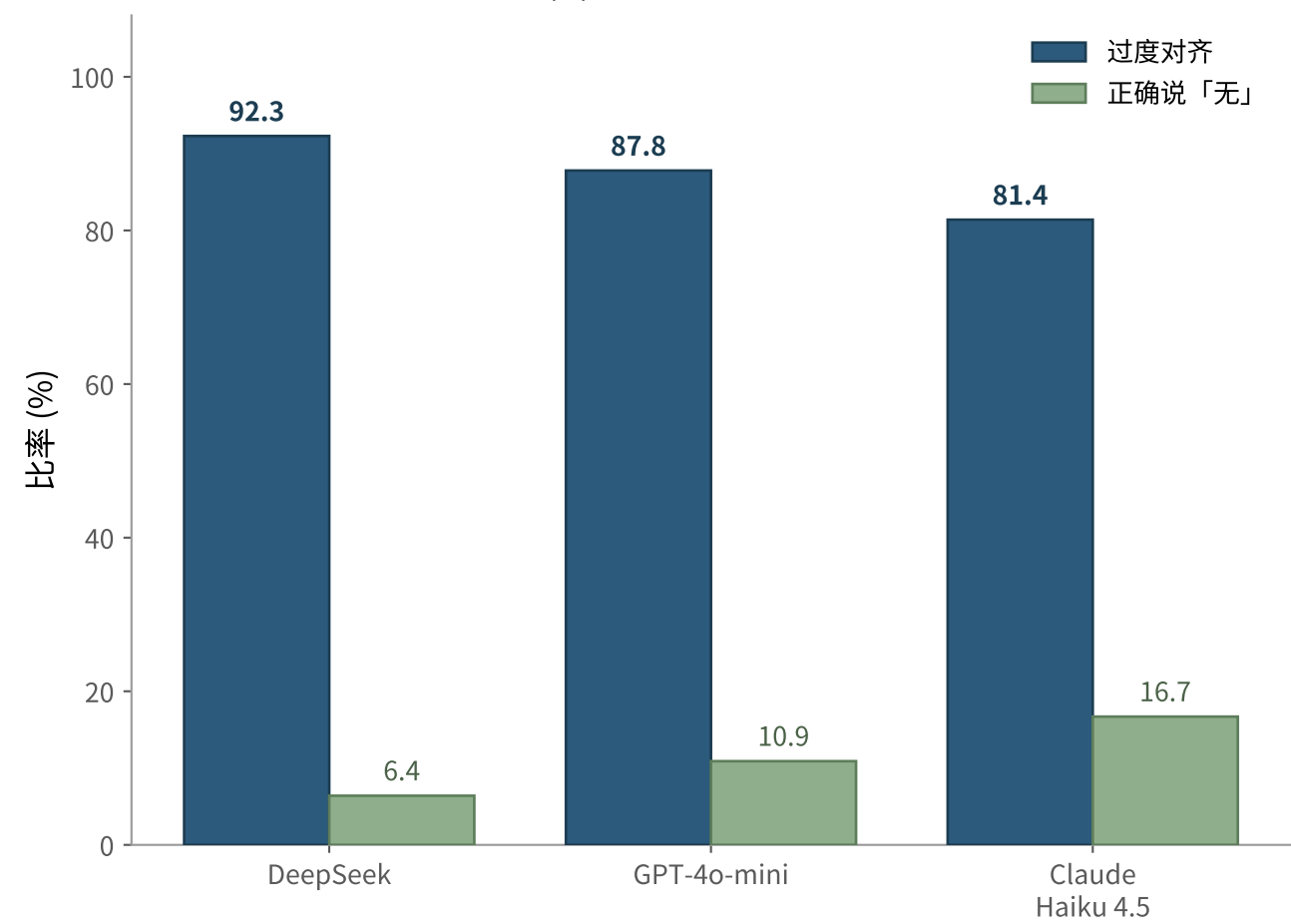
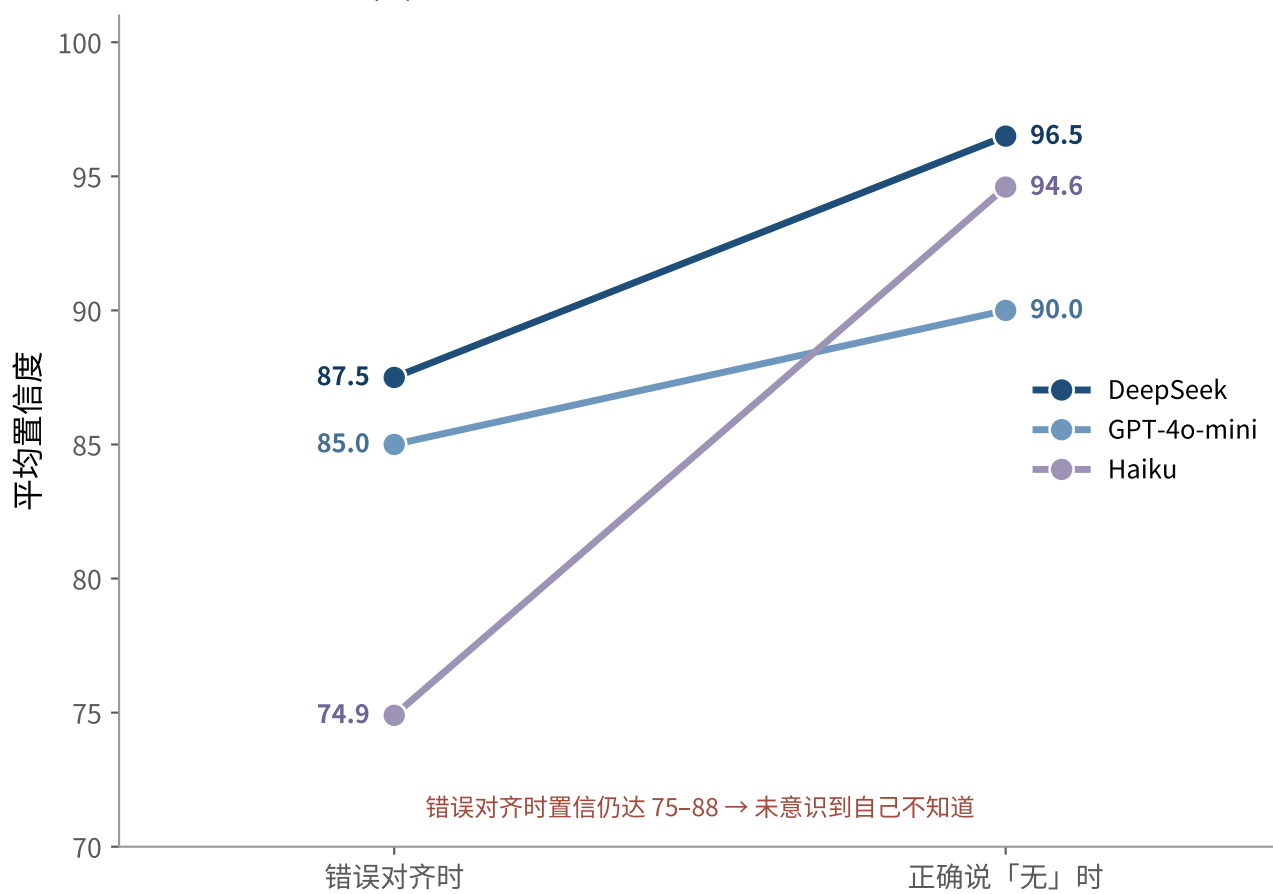


图 A 过度对齐的跨模型机制诊断 (L3 无对应术语, n=156)

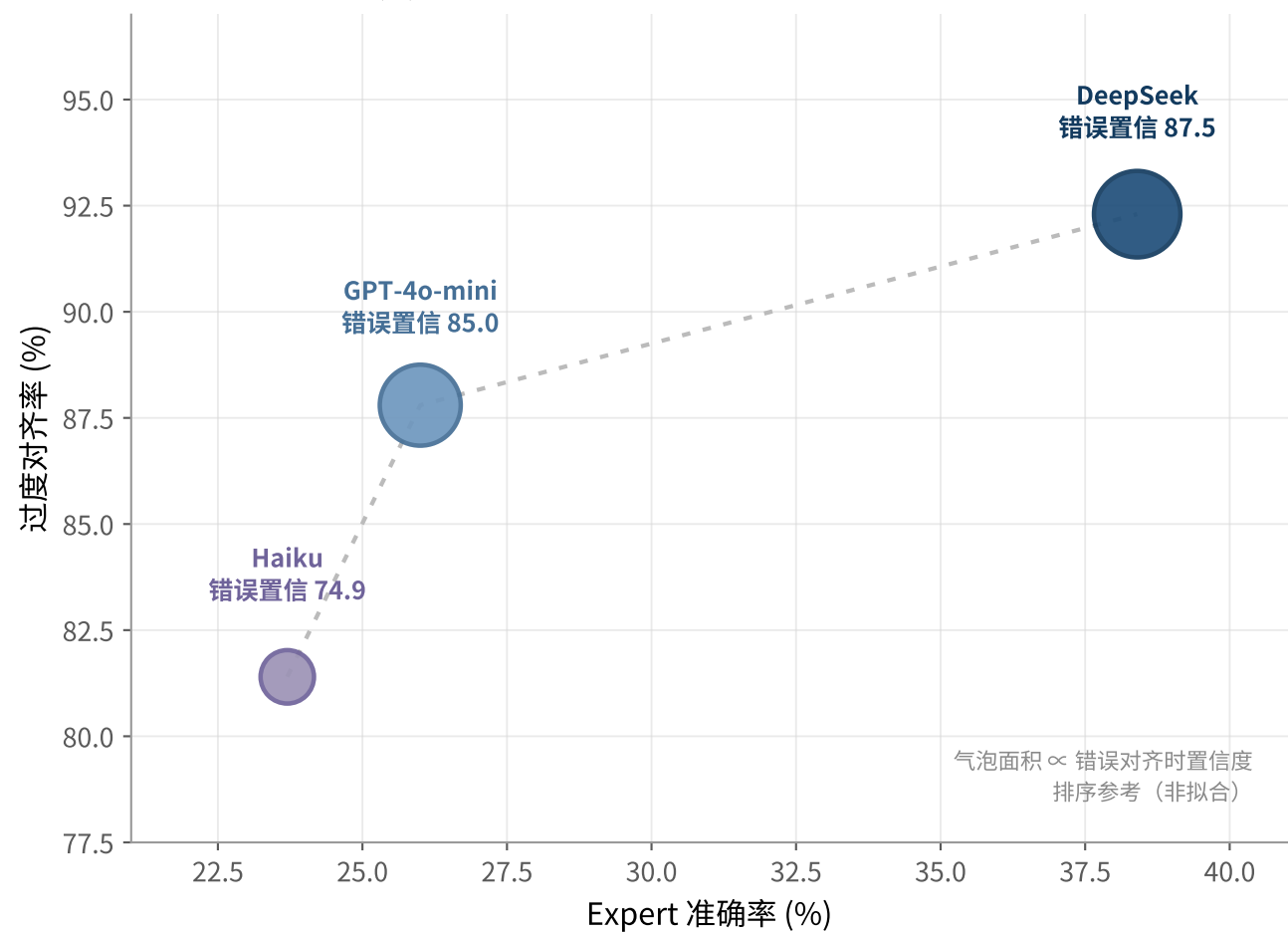
(a) 三模型过度对齐率



(b) 置信度校准：错误对齐 vs 正确说「无」

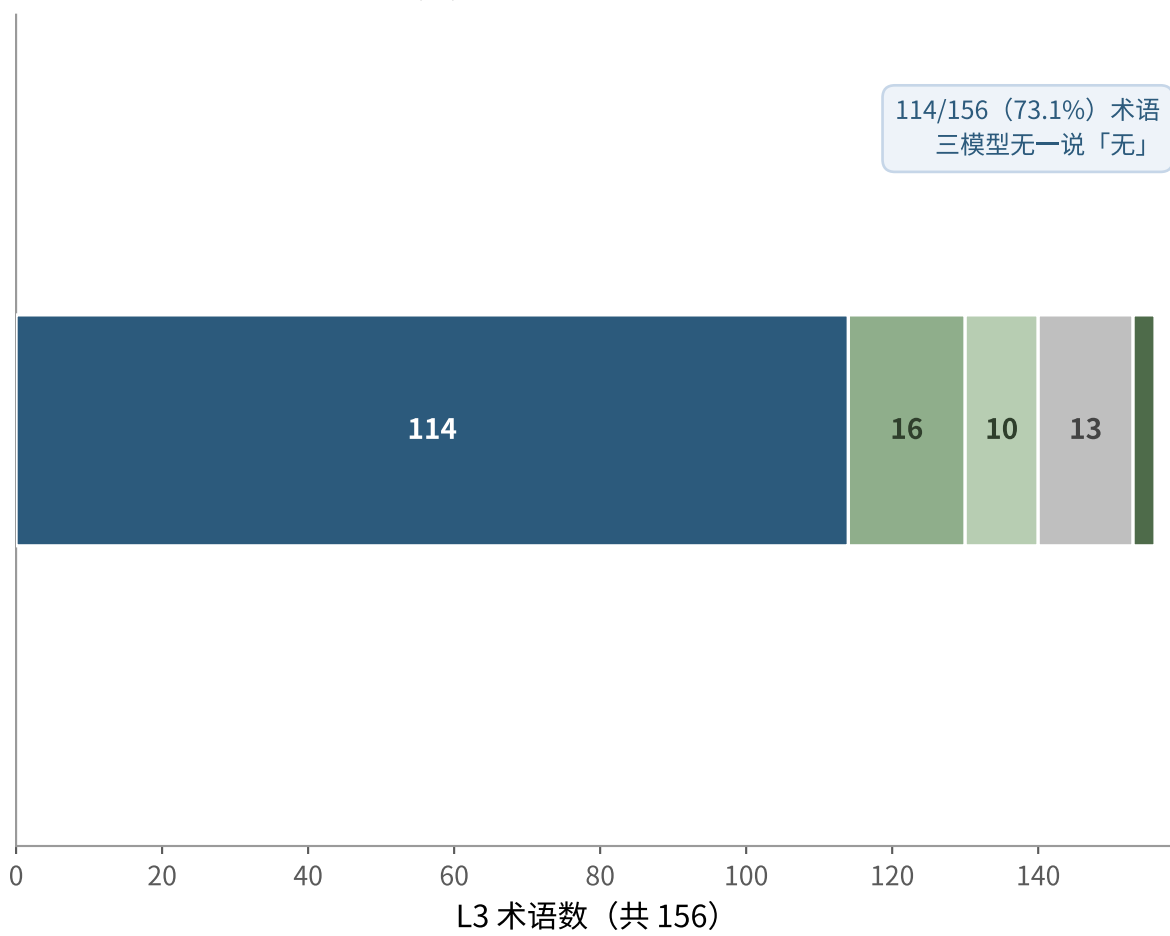


(c) 能力 × 过度对齐 × 过度自信 (描述性)



(c) 描述性观察：能力更强 → 过度对齐更高、且错误时更自信；仅 3 个模型，不作统计推断。

(d) 三模型术语级一致性



三模型全部过度对齐 仅 GPT 说「无」 三模型全部说「无」
仅 Haiku 说「无」 其他混合