量化分析：logit 与 contrib 指标定义 i

为分析各分支对最终决策的影响方向，定义两个指标：

1. logit 指标：

- 为主分类器融合所有分支嵌入后线性层的输出（未经过 Sigmoid）；
- 正值表示模型倾向判为 **AI**，负值表示倾向 **非 AI**；
- 绝对值大小反映决策置信度；
- 可视为样本被判为 AI 的“力度”指标，用于衡量扰动前后决策信号变化。

量化分析：logit 与 contrib 指标定义 i

2. Contribution 指标：

- **Grad×Input，**

$$\mathrm{contrib}_k \approx \sum_{j \in \mathcal{I}_k} \frac{\partial y}{\partial z_j} z_j,$$

  其中 $\mathcal{I}_k$ 表示第 k 个分支在拼接向量中的索引集合。

- $\mathrm{contrib}_k > 0$ 表示该分支对判决结果具有正向推动作用（倾向于判为 AI），$\mathrm{contrib}_k < 0$ 则表示负向作用（倾向于判为非 AI）。

量化分析结果：分支贡献对比

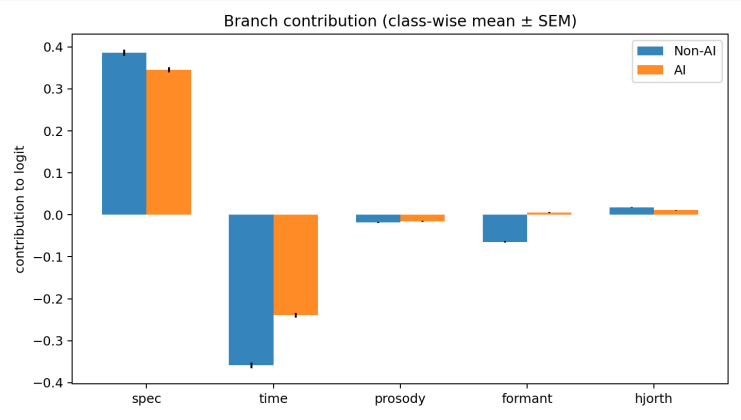

Figure 1.各分支在 AI 与非 AI 样本上的平均贡献值（均值 ± 标准误差）。