PRF 与 Top-k 两幅图的论文级解释（含公式）

本节解释脚本生成的两幅评估图：

- `fig_prf_by_season_k2.png` ：按赛季的 Precision / Recall / F1（把"被淘汰"视为正类）
- `fig_topk_hit_by_season.png` ：按赛季的 Top-k 命中率（k=2 与 k=n_elim）

> 数据来源
>
> - 真实淘汰标签来自 `cleaned_dwts_data_V2.csv` （通过 results 字段解析 `is_eliminated` ）
> - 模型输出的粉丝份额来自 `fan_vote_final_fixed.csv` （字段 `fan_vote_normalized` ）
> - 每周"危险集合（bottom-k）"通过节目规则（分赛季）由评委与粉丝合成得到

> 为什么用这些指标？
>
> - 真实观众投票数/份额不可见（保密），无法对投票份额本身计算 MAE/R^2。
> - 但"每周淘汰结果"可见，因此可以把淘汰建模为一个**分类/检索问题**：
>   - 分类：谁会被淘汰（正类）
>   - 检索：Top-k 风险集合是否覆盖真实淘汰者

---

1. 核心建模口径：每周的"危险集合"如何得到

对任意赛季 s、周 w，设参赛选手集合为 $\mathcal{I}_{s,w}$。

- 评委总分：$S_{i,s,w}$

- 评委百分比：

$$J_{i,s,w} = \frac{S_{i,s,w}}{\sum_{j \in \mathcal{I}_{s,w}} S_{j,s,w}}$$

- 模型估计的粉丝投票份额（每周归一化）：

$$F_{i,s,w} \approx \texttt{fan\_vote\_normalized}$$

节目规则分段（与你的脚本一致）：

1.1 S3–S27：百分比合成（Percent Method）

合成分：

$$C_{i,s,w} = J_{i,s,w} + F_{i,s,w}$$

危险性：C 越小越危险。

因此定义 Top-k 危险集合（bottom-k）：

$$D_{s,w}^{(k)} = \arg\min_k \{C_{i,s,w} : i \in \mathcal{I}_{s,w}\}$$

1.2 S1–S2、S28+：排名合成（Rank Method）

令 $r_{i,s,w}^{J}$ 为评委分数从高到低的名次（1最好），$r_{i,s,w}^{F}$ 为粉丝份额从高到低名次（1最好），则：

$$C_{i,s,w} = r_{i,s,w}^{J} + r_{i,s,w}^{F}$$

危险性：C 越大越危险。

因此 Top-k 危险集合（bottom-k）为：

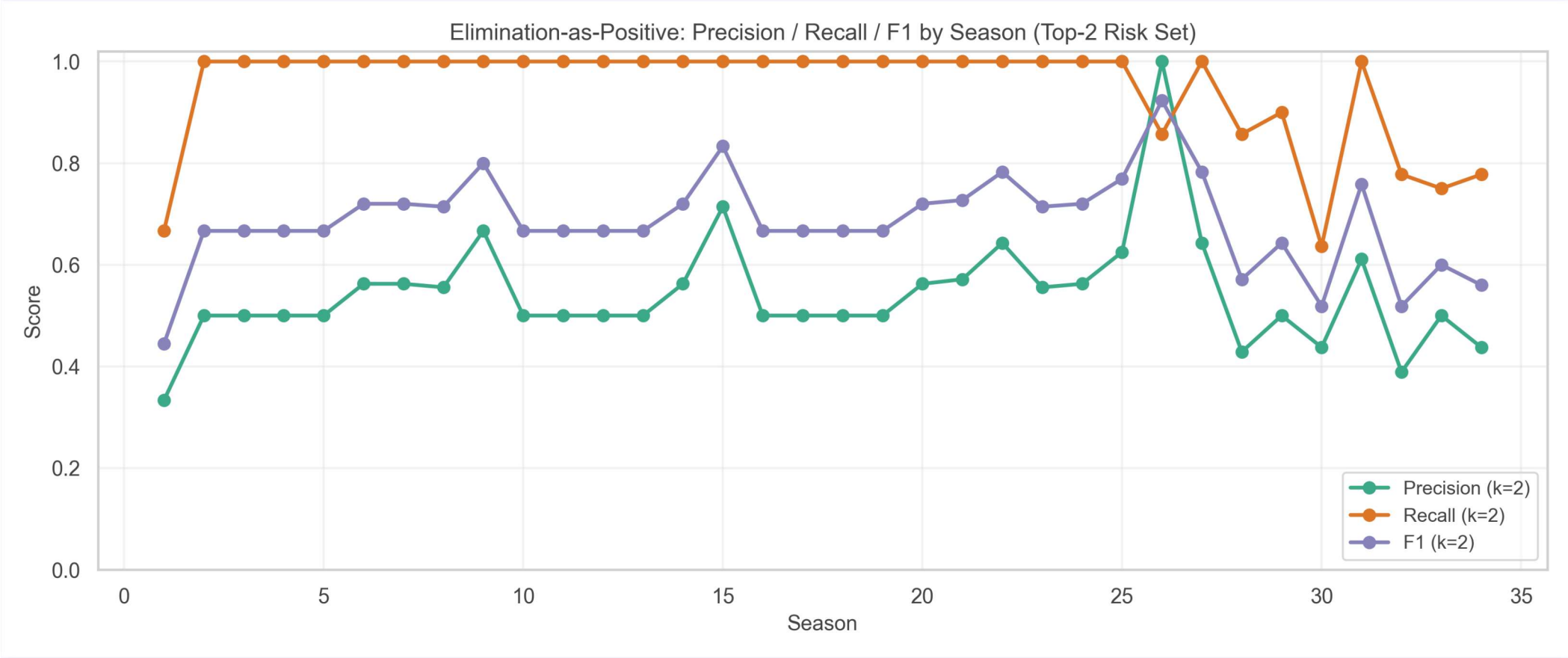
$$D_{s,w}^{(k)} = \arg\max_k \{C_{i,s,w} : i \in \mathcal{I}_{s,w}\}$$

> 注：S28+ 真实淘汰来自 bottom-2（Judges' Save 口径），因此 k=2 的 Top-2 指标具有直接赛制含义。

---

2. 图4：按赛季 Precision / Recall / F1（Top-2 风险集合）

对应图片：



2.1 这张图用来干什么（论文用途）

- **把一致性评估从“周命中/不命中”提升为标准分类指标**，便于与常见机器学习评估接轨。
- **衡量模型对“淘汰正类”的识别能力**：
  - Precision 高：预测的风险人群里“误报（FP）”少
  - Recall 高：真实淘汰者更容易被包含进预测风险集合（漏报（FN）少）
  - F1：在 Precision 与 Recall 之间折中
- 特别适合写在论文“模型验证/鲁棒性评估”小节中，作为 accuracy/hit 的补充。

2.2 指标定义（公式）

对每周 (s, w)：

- 真实淘汰正类集合：

$$E_{s,w} = \{i \in \mathcal{I}_{s,w} \mid i\text{在}(s,w)\text{被淘汰}\}$$

- 预测正类集合（本图采用 $k = 2$）：

$$\hat{E}^{(2)}_{s,w} = D^{(2)}_{s,w}$$

从集合得到 TP/FP/FN：

$$\mathrm{TP}_{s,w} = |E_{s,w} \cap \hat{E}^{(2)}_{s,w}|$$

$$\mathrm{FP}_{s,w} = |\hat{E}^{(2)}_{s,w} \setminus E_{s,w}|$$

$$\mathrm{FN}_{s,w} = |E_{s,w} \setminus \hat{E}^{(2)}_{s,w}|$$

赛季层面采用 **micro-averaging**（先累计再计算，避免每周样本量差异造成偏差）：

$$\mathrm{TP}_s = \sum_{w \in \mathcal{W}_s} \mathrm{TP}_{s,w}, \quad \mathrm{FP}_s = \sum_{w \in \mathcal{W}_s} \mathrm{FP}_{s,w}, \quad \mathrm{FN}_s = \sum_{w \in \mathcal{W}_s} \mathrm{FN}_{s,w}$$

则：

$$\mathrm{Precision}_s = \frac{\mathrm{TP}_s}{\mathrm{TP}_s + \mathrm{FP}_s}$$

$$\mathrm{Recall}_s = \frac{\mathrm{TP}_s}{\mathrm{TP}_s + \mathrm{FN}_s}$$

$$\mathrm{F1}_s = \frac{2\,\mathrm{Precision}_s\,\mathrm{Recall}_s}{\mathrm{Precision}_s + \mathrm{Recall}_s}$$

2.3 读图逻辑（如何写解释）

- 若某赛季 Recall 低：模型经规则合成后，经常**漏掉真实淘汰者**（淘汰者没被列入 bottom-2 风险集合）。这通常发生在：
  - 当周淘汰机制受特殊赛制干扰（双淘汰/复活/非标准流程）
  - 评委与粉丝出现强分歧，导致合成排序不稳定
- 若 Precision 低：模型预测的 bottom-2 风险集合里误报多，说明模型排序存在噪声或合成规则在该季对模型输出更敏感。
- F1 综合衡量上述两点，适合用一句话总结"该季总体淘汰识别质量"。

2.4 可优化说明（让图更像论文图）

- 同图叠加 $k = n_{\text{elim}}$ 的 PRF（尤其在双淘汰周更公平），形成 k=2 与 k=n 的对照。
- 加上置信区间（例如对赛季周数做 bootstrap），让曲线波动更具统计解释力。
- 在背景中标注规则区间（Rank / Percent / Judges' Save），帮助读者理解跨赛季可比性。

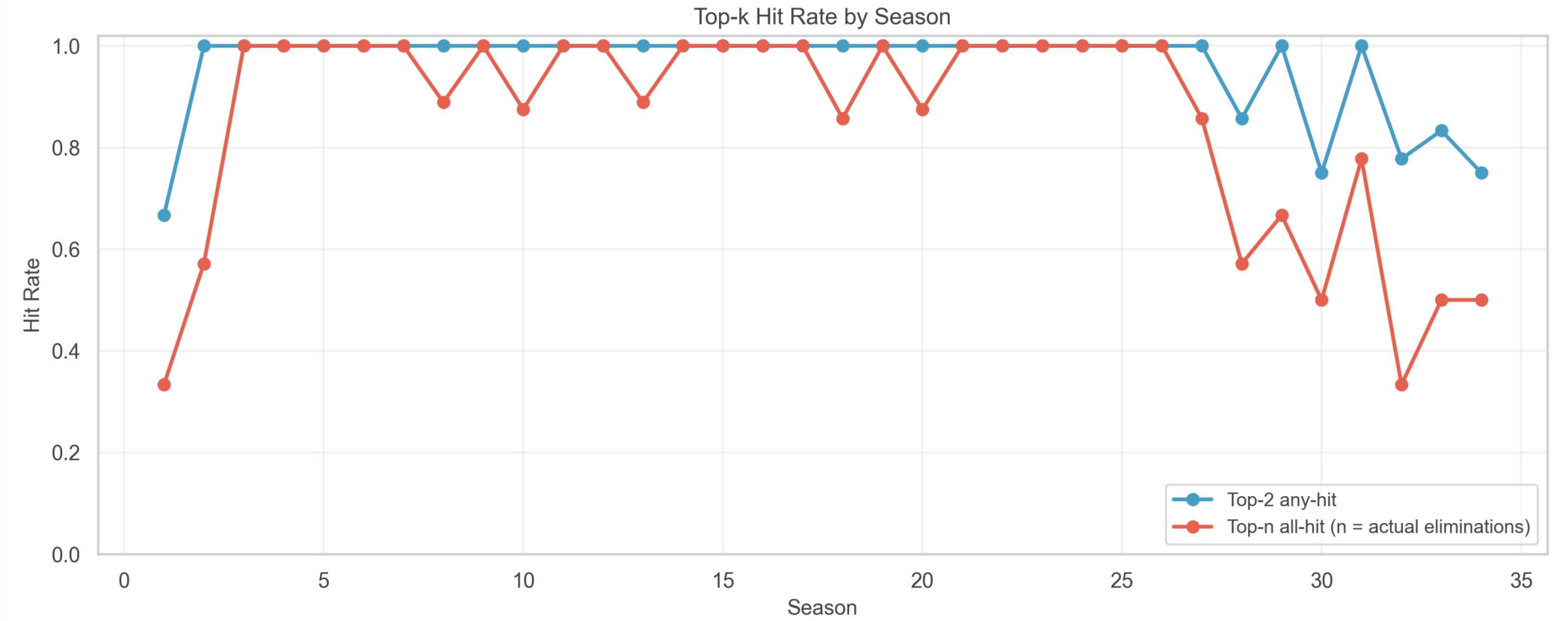
2.5 推荐图注（caption，可直接用）

- "Season-wise micro-averaged Precision/Recall/F1 for elimination detection, where the predicted positive set is the rule-based bottom-2 risk set."

---

3. 图5：按赛季 Top-k 命中率（Top-2 与 Top-n）

对应图片：



3.1 这张图用来干什么（论文用途）

- **把淘汰预测视为"Top-k 检索问题"**：我们不要求精准预测每个人的投票份额，只要求"最危险的 k 人"是否覆盖真实淘汰者。
- **更贴合节目机制的验证方式**：节目最终只淘汰极少数人（1人或2人），因此"危险集合覆盖淘汰者"是最直接的机制一致性验证。
- 适合与题目中的"是否与每周淘汰保持一致"对齐，并可用于比较不同合成方法或不同模型输出。

3.2 指标定义（公式）

设预测危险集合为 $D_{s,w}^{(k)}$，真实淘汰集合为 $E_{s,w}$。

本脚本输出两种 Top-k 命中率：

1. **any-hit（至少命中一个淘汰者）**：

$$\text{Top-}k\_\text{any}(s,w) = \mathbb{I}\left(E_{s,w} \cap D_{s,w}^{(k)} \neq \emptyset\right)$$

2. **all-hit（覆盖全部淘汰者）**：

$$\text{Top-}k\_\text{all}(s,w) = \mathbb{I}\left(E_{s,w} \subseteq D_{s,w}^{(k)}\right)$$

赛季命中率取周平均：

$$\text{HitRate}_s^{\text{any}}(k) = \frac{1}{|\mathcal{W}_s|} \sum_{w \in \mathcal{W}_s} \text{Top-}k\_\text{any}(s,w)$$

$$\text{HitRate}_s^{\text{all}}(k) = \frac{1}{|\mathcal{W}_s|} \sum_{w \in \mathcal{W}_s} \text{Top-}k\_\text{all}(s,w)$$

本图绘制两条典型曲线：

- **Top-2 any-hit**：$\text{HitRate}_s^{\text{any}}(2)$（与 Judges' Save bottom-2 机制强相关）
- **Top-n all-hit**：$\text{HitRate}_s^{\text{all}}(n_{\text{elim}})$（与"完全一致命中"口径一致）

3.3 读图逻辑（论文可写结论）

- Top-2 any-hit 高：说明模型在多数周能把真实淘汰者放入 bottom-2 风险集合，适合解释为"对淘汰风险识别较强"。
- Top-n all-hit 高：说明模型对当周淘汰集合的覆盖更完整（尤其在双淘汰周不只命中其一）。这与"规则一致性 accuracy/hit"高度对应。
- 若两条曲线差异较大：意味着模型经常"命中其一但漏掉另一个"（双淘汰周更常见），可用于讨论模型在多淘汰情境下的局限。

3.4 可优化说明（更强说服力）

- 在图上同时画出 $k = 1, 2, 3$ 的 any-hit 曲线（形成 Top-k 检索性能曲线），体现"k 放宽→命中率上升"的规律。

- 对比不同合成方法（例如 Rank vs Percent）或不同模型版本输出的曲线，用同一图展示改进收益。
- 添加"随机基线"：在每周随机选 k 人作为风险集合，其期望 any-hit 命中率为：

$$\mathbb{E}[\text{any-hit}] \approx 1 - \frac{\binom{N-n_{\text{elim}}}{k}}{\binom{N}{k}}$$

其中 $N = |\mathcal{I}_{s,w}|$。
这样能量化"模型优于随机多少"。

3.5 推荐图注（caption，可直接用）

- "Season-wise Top-k hit rates for elimination retrieval. The Top-2 any-hit rate reflects whether the observed elimination is contained in the rule-based bottom-2 risk set, while the Top-n all-hit rate measures full coverage when n equals the number of actual eliminations."

---

4. 指标文件对应关系（便于论文复现）

运行 `codeMCM/2026/c/Consistency_metrics.py` 后，相关输出为：

- 周度 PRF/Top-k： `prf_topk_metrics_weekly.csv`
- 赛季 PRF/Top-k： `prf_topk_metrics_season.csv`
- 两张图：
  - `fig_prf_by_season_k2.png`

- fig_topk_hit_by_season.png