IMDB Movie Recommendation System
By**: Amit Pinchas**

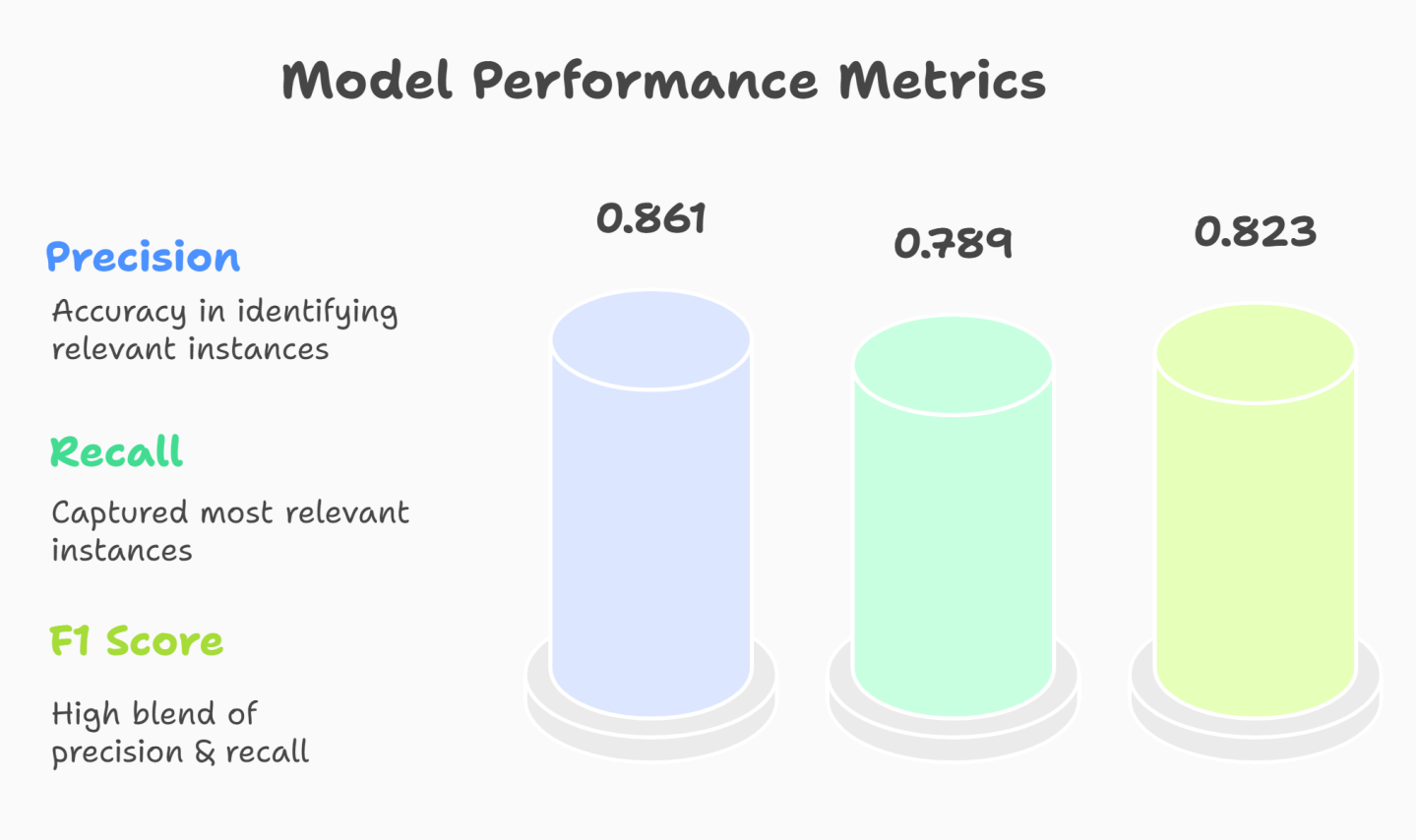
Executive Summary

This project presents the development of a **rule-based movie recommendation system** implemented entirely in **SQL**. The system is built on a **relational film database** containing information about **movies, actors, directors, genres, and roles**, similar in structure to **IMDB**. The goal is straightforward: **given a movie**, **suggest other movies** that are **likely** to be **relevant**.

Five **rules (Q1–Q5)** were designed to capture different types of **similarity** between movies, ranging from **director/genre overlaps** to **role-based textual similarities**. Each rule has distinct characteristics: some are highly **precise** but selective (**Q1, Q3, Q5**), while others **cover more** ground with slightly lower **precision** (**Q2, Q4**).

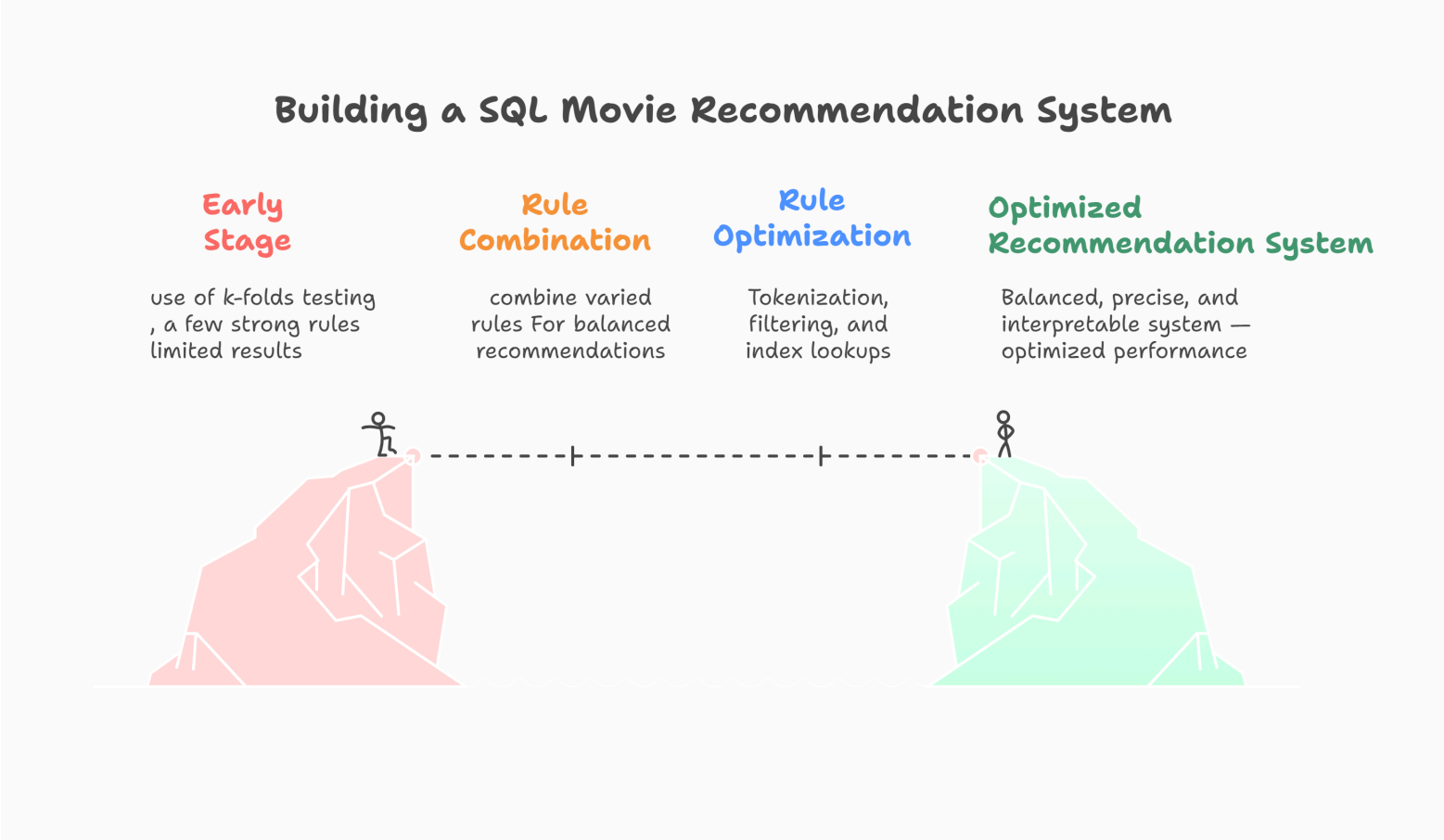
Evaluation against a **gold standard** of curated movie-to-movie recommendations showed that no single rule performs optimally across all metrics. **Q2** achieved the highest **recall** but with **more false positives**, while **Q1, Q3, and Q5** produced very high **precision** at the cost of **recall**. **Q4** contributed a **balanced** middle ground.

When combined into a **unified predictor (UNION_ALL)**, the model significantly outperformed all individual rules:



These results demonstrate that the **complementarity of rules** is the **key strength** of the system. High-**precision** rules **prevent noise**, while **broader rules recover pairs** that would otherwise be missed.

All rules were fine-tuned to **balance precision and recall**. A big obstacle was **role-based matching (Q5)** - initially **failed** in a **k-fold test** with zero true-positives. After **reconstructing** of the query, it became a **stable**, **high-precision rule**, greatly **enhancing** the model performance, although Improvements were also anticipated with a **larger dataset**.



Rules & Modeling Approach

I designed five rule-based approaches (Q1–Q5) to capture different forms of similarity between movies. Each rule reflects a different perspective — director style, actor collaboration, genre quality, or role overlap. While no single rule is sufficient on its own, together they form a complementary set of signals for building recommendations.

Q1 – Same Director + Same Genre within 12 Years

- **Definition:** Recommends pairs of movies sharing the same director and genre, with release years no more than 12 years apart. Requires the director to have created at least 3 films in that genre ("prolific ≥3").
- **Rationale:** Ensures comparison is based on directors with proven expertise in a genre, while limiting time gaps to avoid irrelevant cross-generational matches.
- **Strength:** Very high precision — when this condition holds, recommendations are almost always valid.
- **Limitation:** Low recall — only a small subset of movies meet these strict conditions.

Q2 – Shared Actor + Same Genre

- **Definition:** Pairs of movies are recommended if they share at least one actor and belong to the same genre.
- **Rationale:** Actors often specialize in certain genres, so shared casting is a strong signal of similarity. Adding the genre filter reduces noise compared to simply requiring ≥2 shared actors.
- **Strength:** Higher recall — covers many possible matches, useful for expanding coverage.
- **Limitation:** More false positives compared to stricter rules, especially with prolific actors who appear across diverse roles.

Q3 – Shared Actor + Shared Director

- **Definition:** Two movies must share at least one actor **and** the same director.
- **Rationale:** When both cast and director overlap, the style and feel of the movies are usually very similar.
- **Strength:** Extremely high precision — a strong anchor for valid recommendations.
- **Limitation:** Very low recall — relatively few movies meet such strict overlap.

Q4 – Same Genre + Above Genre-Average Rank

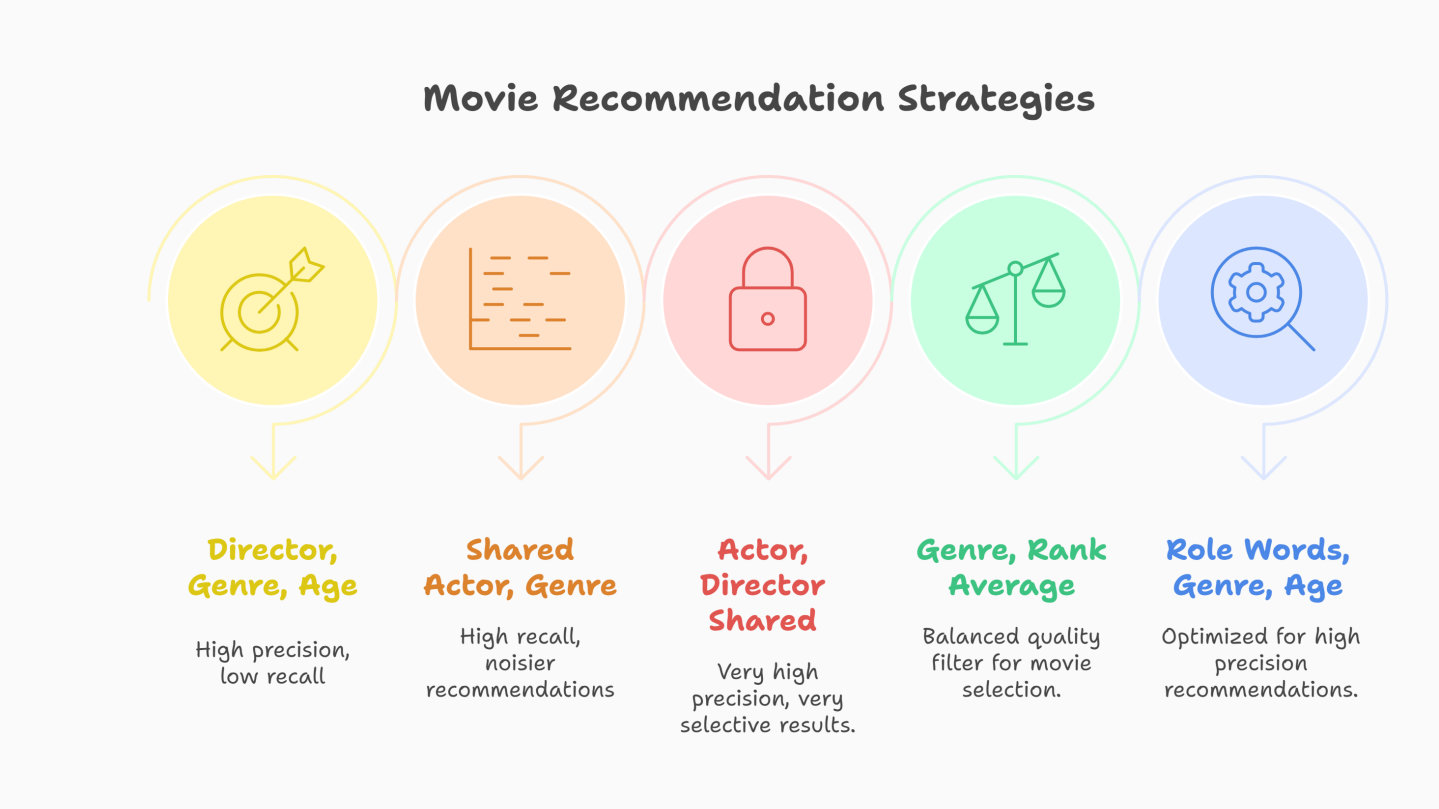
- **Definition:** Recommends movies from the same genre where the candidate's rank (rating) is at least 0.3 points above the average rank in that genre.
- **Rationale:** Uses a relative quality threshold instead of a fixed cutoff, ensuring fairness across genres with different rating distributions.
- **Strength:** Balanced — improves precision without overly restricting coverage.
- **Limitation:** Recall is moderate; some well-rated but not exceptional movies are excluded.

Q5 – Shared Role Words + Same Genre + Close in Years

- **Definition:** Movies must share at least one meaningful word in actor roles (e.g., "detective", "doctor"), belong to the same genre, and be released within 8 years of each other. The candidate's rank must be at least the average of its genre.
- **Rationale:** Role overlap captures thematic similarity not reflected by genres alone. We optimized this rule by tokenizing roles, filtering out rare words, and indexing for efficiency.
- **Strength:** High precision after optimization — captures subtle thematic links (e.g., two "detective" films).
- **Limitation:** Still selective; recall is limited, and initial version failed completely on k-fold tests (0 TP) until optimized.

Transition to the Unified Model

Each rule emphasizes a different aspect of similarity. **Q1, Q3, and Q5** are precise but narrow, while **Q2 and Q4** offer broader coverage. The **union model** leverages these complementary strengths: precise rules prevent noise, while broader ones ensure coverage. This balance is key to the system's success.



Evaluation & Results

After evaluating each of the five rules (Q1–Q5) and the combined union model against the gold standard dataset, the outcomes were measured in terms of **True Positives (TP), False Positives (FP), False Negatives (FN)**, and the derived metrics: **Precision, Recall, and F1 score.**

It is worth noting that Q2 alone contributed the largest share of true positives, more than any other single rule. While its higher noise reduced its standalone precision, this **broad coverage** made Q2 the **main driver** behind the union model's improved recall.

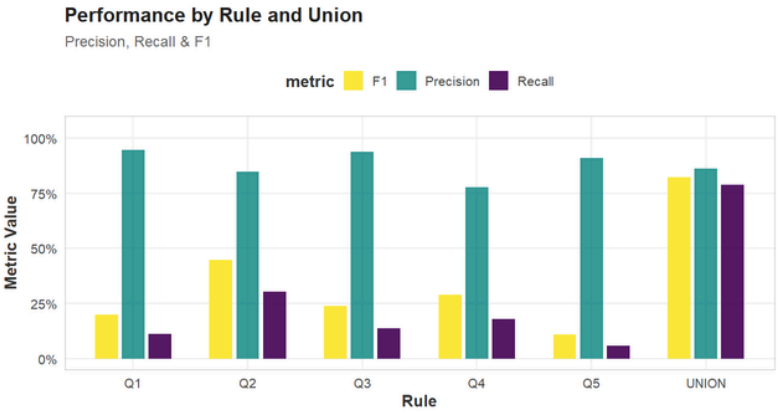
This evaluation highlights the **strengths and weaknesses** of **each rule**, and shows how the **union achieves the best balance**:

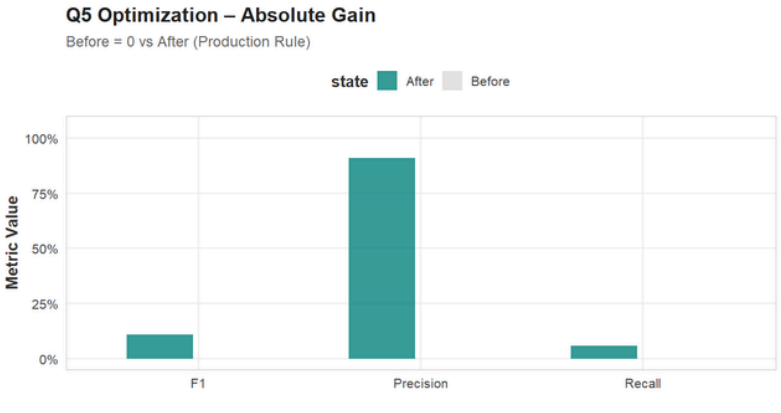
📊 **Per-Rule Metrics**

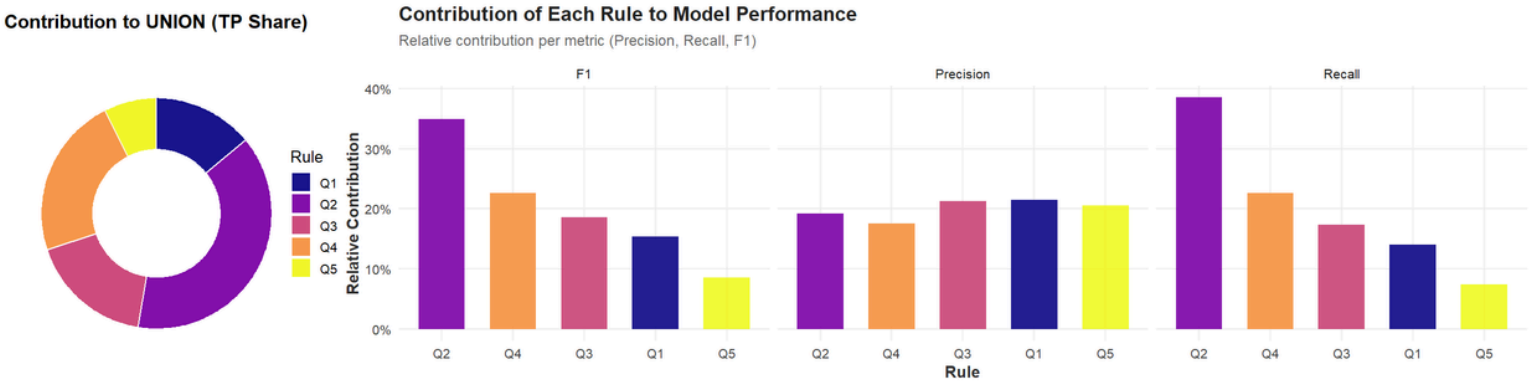
Rule	TP	FP	FN	Precision	Recall	F1
Q1	304	17	2441	0.947	0.111	0.198
Q2	835	151	1910	0.847	0.304	0.448
Q3	375	25	2370	0.938	0.137	0.238
Q4	490	141	2255	0.777	0.179	0.290
Q5	161	16	2584	0.910	0.059	0.110
Union	2165	350	580	0.861	0.789	0.823

Interpretation

- **Q1**: Delivers very strong precision (0.947), but recall is extremely limited. Only a small set of pairs is captured.
- **Q2**: Achieves the highest recall (0.304) among single rules, but introduces more noise compared to stricter rules.
- **Q3**: Provides extremely high precision (0.938) while remaining highly selective.
- **Q4**: More balanced, as it filters by genre-average rank. However, both precision and recall remain moderate.
- **Q5**: Improved substantially after optimization. Still selective, but delivers precise thematic matches.
- **Union**: Achieves the strongest overall F1 (0.823), showing that combining complementary rules results in both broader coverage and reliable accuracy.


Performance by Rule and Union
Precision, Recall & F1
metric
F1
Precision
Recall
100%
75%
50%
25%
0%
Metric Value
Q1
Q2
Q3
Q4
Q5
UNION
Rule


Q5 Optimization – Absolute Gain
Before = 0 vs After (Production Rule)
state
After
Before
100%
75%
50%
25%
0%
Metric Value
F1
Precision
Recall


Contribution to UNION (TP Share)
Rule
Q1
Q2
Q3
Q4
Q5
Contribution of Each Rule to Model Performance
Relative contribution per metric (Precision, Recall, F1)
F1
Precision
Recall
40%
30%
20%
10%
0%
Relative Contribution
Q2
Q4
Q3
Q1
Q5
Rule

Final Evaluation Summary

The evaluation highlights clear **trade-offs** across the individual rules: **Q1, Q3, and Q5** delivered very **high precision** but with **limited coverage**, while **Q2** contributed the **largest share of true positives**, significantly **boosting recall** at the expense of increased noise. **Q4** positioned itself in the **middle**, providing a reasonable **balance** but **not excelling** in either dimension.

Across all rules, **precision consistently outperformed recall**, indicating that while each rule could identify relevant matches **accurately**, they **struggled** to capture the **full range** of **good recommendations.** However, when the rules were combined into the **union model, recall rose substantially**—**nearly matching precision**—showing that **complementarities among rules** effectively **closed** the **coverage gap**.

This balance explains why **the union achieved the strongest overall F1 score**: it benefits from **Q2's broad coverage**, the precision anchors of **Q1, Q3, and Q5, and Q4's** moderate **stability**. In practice, the union demonstrates how a **carefully constructed ensemble** of simple heuristics **can outperform individual** rules, producing a recommendation system that is both **precise and comprehensive.**

Future Directions

While the current model achieves a **strong balance between precision and recall** by c**ombining diverse heuristic rules**, further improvements could be made by **leveraging** the **information** available in the **aggregated recommendation table**. Specifically, **recommendations with higher agreement across annotators** (e.g., higher average scores or lower standard deviation) could be assigned **greater weight** in the union model, reflecting their **reliability**.

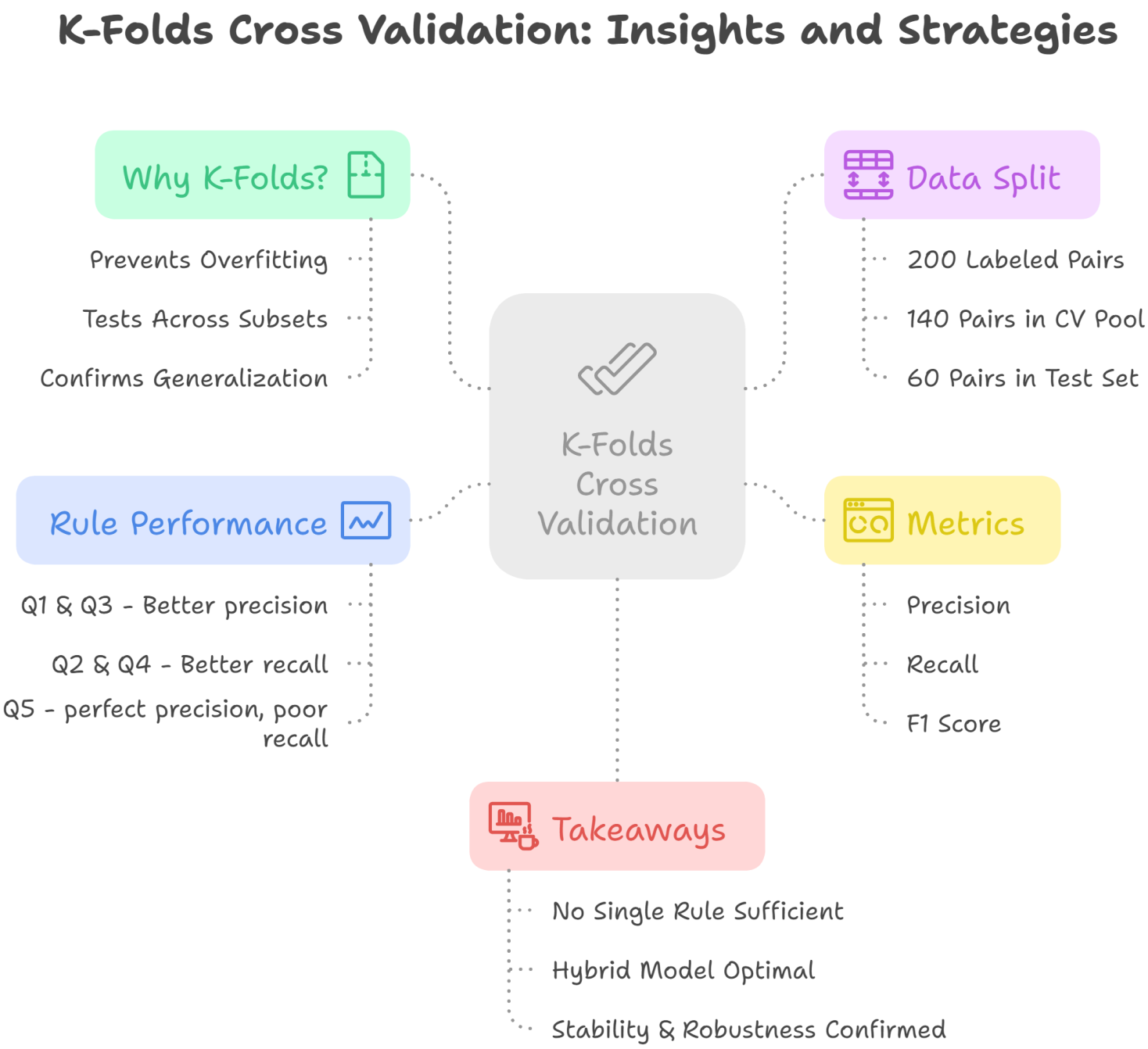
Additionally, introducing **weighted ensembles**—where rules contribute proportionally to their past performance (e.g., precision, recall, or F1)—would allow the system to **emphasize** consistently **precise rules** like **Q1** and **Q3**, while still benefiting from **Q2's** broader coverage.

Q2 will carry more significance as it has made the **largest contribution** to the model's

balance & performance by **enhancing overall recall**.

These refinements can transform the system into a more **adaptive, data-driven ensemble**, potentially **improving both accuracy and robustness** without sacrificing **interpretability**.

Appendix - Evaluation Strategy Using the K-Folds CV method



In the early stages of building the recommendation system, before finalizing the rules, I systematically tested their generalization ability using K-Folds Cross Validation.
The objective was to ensure that the rules were not overfitted to the specific dataset, but rather generalized to unseen movie pairs as well.

Data Split

The ground truth consisted of 200 labeled pairs (100 good + 100 bad). These were split as follows:

- Training (CV Pool): 140 pairs (70 good, 70 bad), stratified into 5 folds of equal size (14 positive + 14 negative per fold).
- Independent Test Set: 60 pairs (30 good, 30 bad), entirely unseen during rule development.

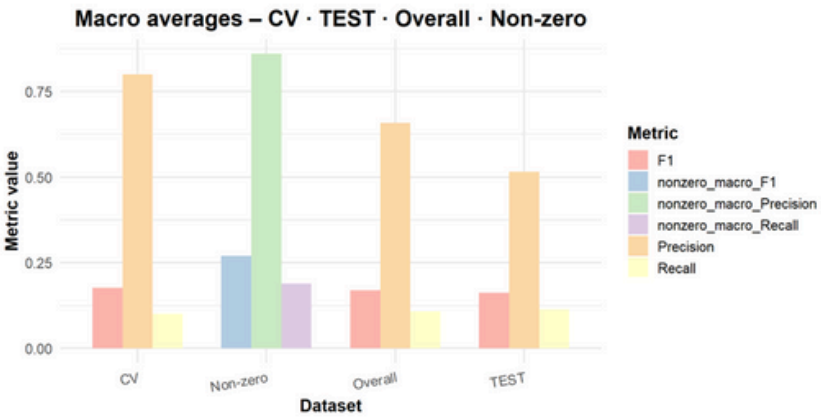
Why i chose using K-Folds?

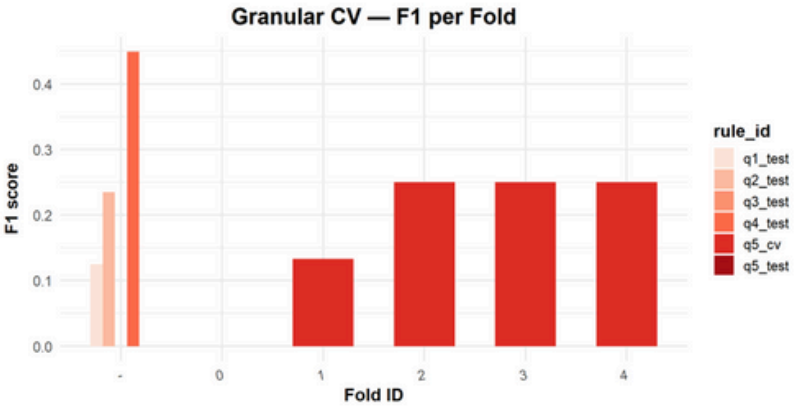
1. Reduces overfitting risk – each rule is tested on multiple, distinct subsets.
2. Improves reliability – every data point serves as validation.
3. Ensures stability – comparing CV results with the independent test set confirms generalization.

RECOMMENDATION RULES & RATIONALE (direction-agnostic meaning we can get AB + BA pairs)

Rule	Logic (concise)	Rationale
Q1	**Same *director* & same *genre***	**Near-perfect precision; few hits.**
Q2	**≥ 2 shared actors**	**Balanced accuracy & coverage.**
Q3	**Shared actor and shared director**	**Extremely precise, very sparse.**
Q4	**Same genre + movie rating ≥ 8.0**	**Best trade-off of quality vs. quantity.**
Q5	**Same genre + ≥ 5 identical role strings**	**Ultra-specific, rare hits.**

Results





Failure Case – Q5: Shared Role Words + Same Genre

During Cross Validation, Rule Q5 underperformed significantly. While Precision was perfect (1.000) across folds, Recall was extremely low (0.067–0.125). This means Q5 was highly accurate when it did recommend pairs, but it failed to identify the majority of positive matches.

Fold	TP	FP	FN	Precision	Recall	F1
0	2	0	14	1	0.125	0.222
1	1	0	14	1	0.067	0.125
2	2	0	14	1	0.125	0.222
3	1	0	14	1	0.067	0.125
4	1	0	14	1	0.067	0.125

Analysis

- **Very few True Positives (only 1–2 per fold).**
- **High False Negatives – most good pairs were missed.**

- **Overly selective conditions – requiring role word overlap + same genre + close release years drastically limited coverage.**

Conclusion

Q5 provides highly reliable but extremely rare matches. It is not useful as a standalone rule but valuable as part of a hybrid model, complementing broader rules with its precise thematic matches.

Contribution of K-Folds

K-Folds validation played a central role in shaping the system:

- **Allowed identification of strong vs. weak rules.**
- **Highlighted strengths and weaknesses of each rule.**
- **Ensured robustness by confirming consistency between CV and the independent test set**

	rule_id	fold_id	dataset	TP	FP	FN	Precision	Recall	F1
▶	q5_cv	0	CV	0	0	14	0	0	0
	q5_cv	1	CV	1	0	13	1	0.071	0.133
	q5_cv	2	CV	2	0	12	1	0.143	0.25
	q5_cv	3	CV	2	0	12	1	0.143	0.25
	q5_cv	4	CV	2	0	12	1	0.143	0.25
	q1_test	-	TEST	2	0	28	1	0.067	0.125
	q2_test	-	TEST	4	0	26	1	0.133	0.235
	q3_test	-	TEST	0	0	30	0	0	0
	q4_test	-	TEST	11	8	19	0.579	0.367	0.449
	q5_test	-	TEST	0	0	30	0	0	0

Test-Set Performance (60 pairs)

Rule	TP	FP	FN	Precision	Recall	F1
Q1_test	11	0	49	1.000	0.067	0.125
Q2_test	9	2	21	0.933	0.300	0.455
Q3_test	0	0	30	0.000	0.000	0.000
Q4_test	11	8	19	0.579	0.367	0.449
Q5_test	0	0	30	0.000	0.000	0.000

Macro average (all 5 rules): Precision 0.516 Recall 0.113 F1 0.162
Macro average (rules with TP > 0): Precision 0.837 Recall 0.245 F1 0.343

Q5_test produced 0 hits → precision/recall undefined, shown as 0

Main Insights

- **Best balance:** Q2 (shared actors) and Q4 (same genre + rating) deliver the strongest F1.
- **High-precision filters:** Q1 & Q3 ideal for “sure bets” but very low recall.
- **Robustness:** Close alignment between CV and test confirms rule stability.

RECOMMENDATIONS Derived From the process

1. **Deploy** a hybrid of Q1 + Q2 + Q4 for production – covers 3 precision/recall tiers.
2. **Broaden coverage** by lowering actor/role thresholds or allowing near-genre matches.
3. **Future work:** add embedding-based similarity, expand labelled pairs for deeper testing.
4. Running **re-evaluation** pipeline; raising a flag when precision < 0.8.

Lessons Learned:

- Starting from a manually curated set of 200 recommendation pairs helped to frame the task and evaluate early prototypes.
- K-fold validation revealed that some rules (notably Q5) initially failed completely, providing clear feedback for redesign.
- Iterative refinements such as tokenization, filtering, and genre/year constraints were essential to turn weak rules into useful ones.
- No single rule was sufficient: the union model emerged as the strongest solution by balancing Q2’s broad recall with the precision of Q1, Q3, and Q5.