# Chapter 3: Advanced Metrics for Evaluating Differential Diagnosis Quality

Evaluating the quality of a differential diagnosis list generated by a Large Language Model (LLM) requires metrics that capture the nuances of clinical reasoning. While established metrics like Top-1 or Top-5 accuracy provide a fundamental assessment of correctness by verifying the presence of the correct diagnosis within the top N suggestions, they may not fully encompass the complexities inherent in constructing a clinically useful differential diagnosis. A well-formed differential list ideally not only includes the correct diagnosis but also ranks diagnoses appropriately by likelihood and ensures that alternative suggestions represent clinically relevant or plausible considerations.

This chapter introduces two advanced, weighted scoring metrics developed within Kernel29: the **Semantic Relationship Score** and the **Severity Matching Score**. These metrics aim to complement traditional accuracy measures by providing a more holistic and clinically meaningful evaluation. They achieve this by acknowledging that diagnoses ranked higher (deemed more likely) should contribute more significantly to the assessment and by factoring in the “distance” (based on semantic relationships or severity levels) between a suggested diagnosis and the true (golden) diagnosis, thereby valuing closer matches more highly than unrelated or significantly mismatched suggestions.

## 1. Complementing Traditional Accuracy Metrics

Traditional metrics such as Top-1 and Top-5 accuracy offer valuable baseline information by treating the differential diagnosis list as a set membership problem, determining if the single most likely diagnosis (Top-1) or any of the top N suggestions (Top-N) match the correct diagnosis. While informative, these metrics primarily focus on the presence or absence of the correct diagnosis within a specified cutoff. They do not inherently capture certain aspects that are often crucial in clinical practice. For instance, Top-N accuracy treats a correct diagnosis identified at rank 1 the same as one found at rank N, potentially overlooking the model’s ability to pinpoint the *most probable* cause. Furthermore, these metrics may not differentiate between a list where alternative diagnoses are plausible clinical considerations (e.g., sharing pathophysiology or belonging to the same disease group) versus a list containing unrelated conditions. They also typically do not assess the alignment between the severity level implied by the suggested diagnoses and the patient’s actual condition severity. Consequently, they evaluate individual list positions for correctness rather than assessing the overall clinical utility and coherence of the differential diagnosis *as a whole*.

To provide a more comprehensive evaluation addressing these facets, Kernel29 employs supplementary weighted scoring methods, detailed below.

## 2. A Holistic Weighted Scoring Approach

To incorporate rank importance and clinical relevance, a weighted scoring methodology is employed. The core idea is to assign a score to each diagnosis in the generated list based on its rank and its “distance” from the golden diagnosis, then compute a weighted average reflecting the overall quality of the list. The general formula structure is:

[ \text{Score} = \frac{\sum\_{N} \left( w\_i \times (\text{D}\_{\text{max}} - \text{D}\_i)2 \right)}{\sum\_^{N} w\_i} ]

Where:

* N represents the number of diagnoses in the LLM’s output list being evaluated (typically 5).
* i is the rank position (index) of a diagnosis within the list, ranging from 1 (most likely) to N.
* w\_i is the weight assigned to rank i, reflecting the principle that higher-ranked (more likely) diagnoses carry more importance. Higher ranks receive higher weights (e.g., using (w\_i = (6-i)/5) for N=5, as shown in Table 1).
* D\_i is the calculated “distance” measuring the mismatch (either semantic or severity-based) between the suggested diagnosis at rank i and the golden diagnosis. A smaller distance indicates a better match.
* D\_max is the maximum possible distance value for the specific metric (semantic or severity). The term (D\_max - D\_i) translates the distance D\_i into a “match score,” rewarding smaller distances (better matches) with higher values. Squaring this term amplifies the reward for very close matches (low D\_i).
* The denominator Σ w\_i represents the sum of all weights applied for the list. This normalization ensures that the final score is an average weighted quality score, allowing for fair comparison even if list lengths (N) or the sum of weights vary (e.g., if N < 5).

Kernel29 utilizes **Linear Weights** for w\_i, prioritizing higher ranks as detailed in Table 1:

**Table 1: Linear Rank Weights (N=5)** | Rank (i) | Weight (w\_i = (6-i)/5) | | :—–– | :——————— | | 1 | 1.0 | | 2 | 0.8 | | 3 | 0.6 | | 4 | 0.4 | | 5 | 0.2 | | **Sum** | **3.0** |

This linear weighting scheme ensures that a correct or highly relevant diagnosis at Rank 1 contributes substantially more to the final score than the same diagnosis placed at Rank 5, reflecting clinical prioritization.

## 3. LLM-as-Judge for Classification

To systematically obtain the semantic relationship categories (Table 2) and severity levels (Table 3) needed for the scoring calculations, an automated approach using a Large Language Model as a classifier, often termed “LLM-as-judge”, was employed. Specifically, Llama 3.1 70B was utilized.

* **Severity Classification:** Llama 3.1 70B was used to classify the severity level for:
  + *Every* diagnosis predicted by the target LLM.
  + The *golden* diagnosis for a case, *but only if* the severity level was not already provided in the source dataset. In both scenarios, the LLM was prompted to classify the severity into one of the categories: mild, moderate, severe, critical, or rare based on the diagnosis name (a multi-class classification task).
* **Semantic Relationship Classification:** For *each pair* consisting of the golden diagnosis and a predicted diagnosis from the target LLM’s list, Llama 3.1 70B was prompted to classify the semantic relationship *between* them. It selected one category from Table 2: Exact Synonym, Broad Synonym, Exact Disease Group, Broad Disease Group, or Not Related.

This automated classification provided the necessary categorical inputs (D\_i for semantics, and the severity values used to calculate D\_i for severity) required by the weighted scoring formulas (Equation 1). The reliability of these classifications depends on the capability of the judge LLM (Llama 3.1 70B) and the specific prompts used. As noted in the Appendix, this implementation serves as an initial example, and further refinement might be necessary for robust, production-level evaluation.

## 4. Combined Score Calculation Examples

To illustrate how these weighted scores capture different facets of differential diagnosis quality, this section examines three illustrative cases from the c3opus model evaluated on the PUMCH\_ADM dataset (ramedis bench) using the dxgpt\_main prompt. Both the Semantic Relationship Score and the Severity Matching Score will be calculated and discussed for each case.

### 4.1 Semantic Relationship Score Definition

The Semantic Relationship Score quantifies the clinical relatedness between the suggested diagnoses and the golden diagnosis. The distance D\_i for this score is **directly assigned** based on predefined clinical relationships mapped to discrete distance values, as shown in Table 2. (The method for determining these relationships is detailed in Section 3.3).

**Table 2: Semantic Relationship Distances** | Semantic Relationship | Distance (D\_i) | Interpretation | | :––––––––––– | :————— | :—————————–– | | Exact Synonym | 1 | Perfect match | | Broad Synonym | 2 | Closely related term | | Exact Disease Group | 3 | Same specific disease category | | Broad Disease Group | 4 | Related broader disease category | | Not Related | 5 | Unrelated condition |

The maximum distance (D\_max) for this score is 5, and the linear weights from Table 1 are applied.

### 4.2 Severity Matching Score Definition

The Severity Matching Score assesses the alignment between the severity level of the suggested diagnoses and the severity level of the golden diagnosis. Unlike the direct assignment used for semantic distance, the severity distance involves calculation based on numerical values.

* **Severity Values**: First, clinical severity levels are mapped to ordinal numerical values as shown in Table 3. Both the severity of the *golden diagnosis* and the severity of *each predicted diagnosis* were determined using an automated classification approach described in Section 3.3.
* **Table 3: Severity Level Values** | Severity Level | Value | | :———–– | :–– | | mild | 1 | | moderate | 2 | | severe | 3 | | critical | 4 | | rare | 5 |
* **Severity Distance (D\_i)**: After converting the severity labels (for both the golden and predicted diagnosis at rank i) to their numerical values using Table 3, the distance D\_i is calculated based on their absolute difference: [ D\_i = 1 + |\text{golden\_severity\_value} - \text{predicted\_severity\_value}| ] This formula ensures the distance reflects the magnitude of the severity mismatch:
  + An exact match (e.g., rare vs rare, value 5 vs 5) results in the minimum distance: (D\_i = 1 + |5-5| = 1).
  + The maximum mismatch (e.g., rare vs mild, value 5 vs 1) results in the maximum distance: (D\_i = 1 + |5-1| = 5).
* **Maximum Distance (D\_max)**: Derived from the calculation method, the maximum possible severity distance is 5. This value is used for D\_max in the general scoring formula (Equation 1) when calculating the Severity Score.
* **Weights (w\_i)**: The same linear weights defined in Table 1 ([1.0, 0.8, 0.6, 0.4, 0.2] for N=5) are applied.

### 4.3 Calculation Examples

Now, let’s apply these definitions to our three cases, selected to represent best, medium, and worst overall performance based on combined scores.

#### 4.3.1 Case 1: Best Performance (Patient ID 31)

* **Golden Diagnosis:** “Myasthenia gravis” (Severity: rare = 5)
* **LLM Output (N=3):**

| Rank | Predicted Diagnosis | Severity | Semantic Relationship |
| --- | --- | --- | --- |
| 1 | Myasthenia Gravis | rare | Exact Synonym |
| 2 | Lambert-Eaton Myasthenic Syndrome (LEMS) | rare | Exact Disease Group |
| 3 | Botulism | severe | Broad Disease Group |

* **Semantic Score Calculation (Adjusted for N=3, Sum Weights = 2.4):**

| Rank | Weight (w\_i) | Semantic Relationship | Sem. Distance (D\_i) | Score Calc: w\_i\*(5-D\_i)² | Contribution |
| --- | --- | --- | --- | --- | --- |
| 1 | 1.0 | Exact Synonym | 1 | 1.0 \* (5-1)² = 16.0 | 16.0 |
| 2 | 0.8 | Exact Disease Group | 3 | 0.8 \* (5-3)² = 3.2 | 3.2 |
| 3 | 0.6 | Broad Disease Group | 4 | 0.6 \* (5-4)² = 0.6 | 0.6 |
|  | **Sum = 2.4** |  |  | **Numerator:** | **19.8** |

**Semantic Score = 19.8 / 2.4 = 8.25**

* **Severity Score Calculation (Adjusted for N=3, Sum Weights = 2.4):** (Golden Severity = 5)

| Rank | Weight (w\_i) | Predicted Severity | Sev. Distance (D\_i = 1 + |5 - pred|) | Score Calc: w\_i\*(5-D\_i)² | Contribution |
| --- | --- | --- | --- | --- | --- |
| 1 | 1.0 | rare (5) | 1 | 1.0 \* (5-1)² = 16.0 | 16.0 |
| 2 | 0.8 | rare (5) | 1 | 0.8 \* (5-1)² = 12.8 | 12.8 |
| 3 | 0.6 | severe (3) | 3 | 0.6 \* (5-3)² = 2.4 | 2.4 |
|  | **Sum = 2.4** |  |  | **Numerator:** | **31.2** |

**Severity Score = 31.2 / 2.4 = 13.0**

#### 4.3.2 Case 2: Medium Performance (Patient ID 54)

* **Golden Diagnosis:** “Myofibrillar myopathy ZASP-related” (Severity: rare = 5)
* **LLM Output:**

| Rank | Predicted Diagnosis | Severity | Semantic Relationship |
| --- | --- | --- | --- |
| 1 | Myofibrillar Myopathy (MFM) | rare | Broad Synonym |
| 2 | Desminopathy | rare | Broad Synonym |
| 3 | Limb-Girdle Muscular Dystrophy (LGMD) | rare | Exact Disease Group |
| 4 | Inclusion Body Myositis (IBM) | severe | Broad Disease Group |
| 5 | Pompe Disease | rare | Broad Disease Group |

* **Semantic Score Calculation:**

| Rank | Weight (w\_i) | Semantic Relationship | Sem. Distance (D\_i) | Score Calc: w\_i\*(5-D\_i)² | Contribution |
| --- | --- | --- | --- | --- | --- |
| 1 | 1.0 | Broad Synonym | 2 | 1.0 \* (5-2)² = 9.0 | 9.0 |
| 2 | 0.8 | Broad Synonym | 2 | 0.8 \* (5-2)² = 7.2 | 7.2 |
| 3 | 0.6 | Exact Disease Group | 3 | 0.6 \* (5-3)² = 2.4 | 2.4 |
| 4 | 0.4 | Broad Disease Group | 4 | 0.4 \* (5-4)² = 0.4 | 0.4 |
| 5 | 0.2 | Broad Disease Group | 4 | 0.2 \* (5-4)² = 0.2 | 0.2 |
|  | **Sum = 3.0** |  |  | **Numerator:** | **19.2** |

**Semantic Score = 19.2 / 3.0 = 6.4**

* **Severity Score Calculation:** (Golden Severity = 5)

| Rank | Weight (w\_i) | Predicted Severity | Sev. Distance (D\_i = 1 + |5 - pred|) | Score Calc: w\_i\*(5-D\_i)² | Contribution |
| --- | --- | --- | --- | --- | --- |
| 1 | 1.0 | rare (5) | 1 | 1.0 \* (5-1)² = 16.0 | 16.0 |
| 2 | 0.8 | rare (5) | 1 | 0.8 \* (5-1)² = 12.8 | 12.8 |
| 3 | 0.6 | rare (5) | 1 | 0.6 \* (5-1)² = 9.6 | 9.6 |
| 4 | 0.4 | severe (3) | 3 | 0.4 \* (5-3)² = 1.6 | 1.6 |
| 5 | 0.2 | rare (5) | 1 | 0.2 \* (5-1)² = 3.2 | 3.2 |
|  | **Sum = 3.0** |  |  | **Numerator:** | **43.2** |

**Severity Score = 43.2 / 3.0 = 14.4**

#### 4.3.3 Case 3: Worst Performance (Patient ID 20)

* **Golden Diagnosis:** “Brugada syndrome” (Severity: rare = 5)
* **LLM Output:**

| Rank | Predicted Diagnosis | Severity | Semantic Relationship |
| --- | --- | --- | --- |
| 1 | Benign Prostatic Hyperplasia (BPH) | mild | Not Related |
| 2 | Myasthenia Gravis | rare | Not Related |
| 3 | Rhabdomyolysis | severe | Not Related |
| 4 | Syncope (Fainting) | mild | Broad Disease Group |
| 5 | Acute Coronary Syndrome (ACS) | critical | Broad Disease Group |

* **Semantic Score Calculation:**

| Rank | Weight (w\_i) | Semantic Relationship | Sem. Distance (D\_i) | Score Calc: w\_i\*(5-D\_i)² | Contribution |
| --- | --- | --- | --- | --- | --- |
| 1 | 1.0 | Not Related | 5 | 1.0 \* (5-5)² = 0.0 | 0.0 |
| 2 | 0.8 | Not Related | 5 | 0.8 \* (5-5)² = 0.0 | 0.0 |
| 3 | 0.6 | Not Related | 5 | 0.6 \* (5-5)² = 0.0 | 0.0 |
| 4 | 0.4 | Broad Disease Group | 4 | 0.4 \* (5-4)² = 0.4 | 0.4 |
| 5 | 0.2 | Broad Disease Group | 4 | 0.2 \* (5-4)² = 0.2 | 0.2 |
|  | **Sum = 3.0** |  |  | **Numerator:** | **0.6** |

**Semantic Score = 0.6 / 3.0 = 0.2**

* **Severity Score Calculation:** (Golden Severity = 5)

| Rank | Weight (w\_i) | Predicted Severity | Sev. Distance (D\_i = 1 + |5 - pred|) | Score Calc: w\_i\*(5-D\_i)² | Contribution |
| --- | --- | --- | --- | --- | --- |
| 1 | 1.0 | mild (1) | 5 | 1.0 \* (5-5)² = 0.0 | 0.0 |
| 2 | 0.8 | rare (5) | 1 | 0.8 \* (5-1)² = 12.8 | 12.8 |
| 3 | 0.6 | severe (3) | 3 | 0.6 \* (5-3)² = 2.4 | 2.4 |
| 4 | 0.4 | mild (1) | 5 | 0.4 \* (5-5)² = 0.0 | 0.0 |
| 5 | 0.2 | critical (4) | 2 | 0.2 \* (5-2)² = 1.8 | 1.8 |
|  | **Sum = 3.0** |  |  | **Numerator:** | **17.0** |

**Severity Score = 17.0 / 3.0 = 5.67**

## 5. Interpretation of Combined Scores

Presenting the Semantic and Severity scores side-by-side allows for a more holistic interpretation of the LLM’s performance on individual cases.

* **Case 1 (Best Performance - Patient ID 31):** This case demonstrated strong performance, scoring well on both Semantic (8.25) and Severity (13.0) metrics. The model identified the “Exact Synonym” at rank 1 and included relevant related diagnoses (“Exact/Broad Disease Group”). Severity matching was also strong, featuring perfect matches in the top ranks. It is noteworthy that the scores for this case were calculated using N=3 (sum of weights = 2.4), as only three diagnoses were provided in the LLM output. By normalizing the total weighted ‘goodness’ score by the sum of weights ((\sum w\_i)), the final score reflects the average quality across the provided ranks. This normalization inherently penalizes longer lists that might include lower-quality items, as the denominator ((\sum w\_i)) grows with list length (N) while the numerator might not increase proportionally if poor suggestions are added. Consequently, this case represents a high-quality differential diagnosis list according to these metrics.
* **Case 2 (Medium Performance - Patient ID 54):** This case achieved a good Semantic score (6.4) and a very good Severity score (14.4). Although the exact diagnosis was not listed, the top ranks were occupied by highly relevant “Broad Synonyms” and “Exact Disease Groups.” Severity matching was excellent, with only a minor discrepancy at rank 4. This output represents a clinically useful list, offering closely related alternatives with appropriate severity alignment.
* **Case 3 (Worst Performance - Patient ID 20):** This case exhibited poor performance, scoring extremely low on Semantics (0.2) and poorly on Severity (5.67). The model failed to include the correct diagnosis (“Brugada syndrome”) or any closely related synonyms within the top 5. The list predominantly consisted of unrelated conditions, and the severity matching was weak, showing significant mismatches (e.g., mild/critical suggested vs. rare actual). This outcome signifies a low-quality list offering limited diagnostic value.

These examples demonstrate how the combination of Semantic and Severity scores provides a nuanced picture of the LLM’s diagnostic reasoning capabilities that extends beyond simple accuracy. The normalization method, dividing by the sum of weights ((\sum w\_i)), ensures the final score represents an average weighted quality. This approach effectively penalizes the inclusion of low-quality items in longer lists and facilitates fair comparisons between lists of potentially different lengths (N).

## 6. Advantages of Weighted Scoring

The Semantic and Severity scores offer significant advantages as complementary metrics to traditional Top-N accuracy for evaluating differential diagnoses. They provide a **holistic evaluation** by assessing the entire list, rather than solely focusing on the presence or absence of the correct diagnosis. Their **rank awareness** explicitly rewards models for placing more likely diagnoses higher in the list. Furthermore, they incorporate **clinical relevance** by considering semantic or severity “closeness,” thereby differentiating between near misses and completely irrelevant suggestions. This results in a continuous score that allows for more **nuanced comparisons** between different models or prompting strategies. Ultimately, the emphasis on ranking and relevance aims for better **alignment with clinical practice**, reflecting how clinicians typically formulate and utilize differential diagnoses.

By employing these advanced metrics, Kernel29 facilitates a more robust and clinically meaningful assessment of LLM performance in the complex task of medical diagnosis. These scores allow researchers to move beyond simple accuracy measures and gain deeper insights into how well models capture the structural and clinical plausibility essential for a useful differential diagnosis.

## 7. Aggregating Scores for Overall Evaluation

While calculating scores for individual clinical cases provides valuable insights, evaluating the overall performance of a specific model-prompt combination across an entire dataset necessitates a method for score aggregation. Kernel29 employs a multi-step process involving score rescaling followed by weighted aggregation to achieve this.

### 7.1 Rescaling Individual Scores

As a preliminary step before aggregation, the individual Semantic and Severity scores, which initially range from 0 to a theoretical maximum of 16 (calculated as (w\_1 \times (D\_{max} - D\_{min})2 = 1.0 \times (5-1)2 = 16)), are rescaled to a standardized range of [-1, 1]. This normalization facilitates comparison across metrics and is a prerequisite for the subsequent weighted aggregation and Cartesian visualization. The rescaling transformation is defined as:

[ \text{score}*{\text{rescaled}} = \frac{\text{score}*{\text{original}} \times 2.0}{\text{score}\_{\text{max}}} - 1.0 ]

Where score\_original is the calculated Semantic or Severity score (0-16) for a single case, and score\_max is the theoretical maximum score (16.0). Under this transformation, an original score of 16.0 maps to +1.0 (best), 8.0 maps to 0.0 (neutral), and 0.0 maps to -1.0 (worst).

### 7.2 Weighted Aggregation

A simple arithmetic mean of rescaled scores might mask clinically significant variations in performance. A model could perform well on average yet occasionally generate severely misleading differential diagnoses (very low scores) on challenging cases. In clinical practice, such significant errors, which could lead to diagnostic delays or mismanagement, are often more critical than minor inaccuracies on straightforward cases. To better reflect this clinical reality and provide a measure more sensitive to potentially harmful outputs, Kernel29 employs a *weighted aggregation* strategy. This approach assigns higher weights to cases where the model performed poorly (lower rescaled scores), ensuring that instances of significant clinical misalignment disproportionately influence the final aggregate score. This yields a summary statistic that is more indicative of the model’s reliability and robustness against critical failures.

The aggregation process involves two main steps:

1. **Calculate Case Weight:** A weight ((W\_{\text{case}})) is determined for each individual case based on its rescaled score using a reversed logistic sigmoid function. This function is chosen for its smooth transition properties. The general form, allowing for tunable sensitivity, is: [ W\_{\text{case}} = \frac{1.0}{1.0 + e^{k \times (\text{score}\_{\text{rescaled}} - x\_0)}} ]
   * score\_rescaled: The input rescaled score for the case (-1 to 1).
   * k: The **steepness** parameter. Higher values create a sharper transition in weight assignment around the midpoint.
   * x\_0: The **midpoint** parameter, representing the rescaled score at which the weight equals 0.5. Shifting x0 adjusts the threshold for what constitutes a “neutral” score in terms of weighting.
   * The function is effectively **reversed** (due to the positive exponent term k \times ... combined with the reciprocal structure) such that it assigns higher weights to scores *below* the midpoint x0 and lower weights to scores *above* it.
   * **Kernel29 Default:** The standard implementation uses x\_0 = 0, centering the neutral weight point (0.5) precisely at a rescaled score of 0.
2. **Calculate Weighted Average:** The final aggregated score for a specific model-prompt combination over the dataset is computed as the weighted average of the individual rescaled scores: [ \text{Score}*{\text{aggregated}} = \frac{\sum (W*{\text{case}} \times \text{score}*{\text{rescaled}})}{\sum W*{\text{case}}} ] This calculation yields a single aggregated, rescaled Semantic score and a single aggregated, rescaled Severity score (both ranging from -1 to 1) for each model-prompt combination evaluated.

#### 7.2.1 Aggregation Examples under Different Difficulty Settings

To illustrate the impact of the weighting parameters (k and x0) on the final aggregated score under varying performance distributions, Table 4 analyzes three hypothetical scenarios using different weighting configurations, termed “difficulty settings”:

* **Easy Setting:** k=1, x0=0.3 (Gentle slope, midpoint shifted right – less emphasis on penalizing scores slightly below 0)
* **Medium Setting:** k=2, x0=0 (Moderate slope, midpoint at zero – moderate penalty for scores below 0)
* **Hard Setting:** k=3, x0=0 (Sharp slope, midpoint at zero – strong emphasis on penalizing scores below 0)

**Table 4: Weighted Aggregation Examples under Different Settings** | Case ID | Score (s) | Weight Calc (Easy: k=1, x0=0.3) | Weight (Easy) | Contrib. (Easy) | Weight Calc (Med: k=2, x0=0) | Weight (Med) | Contrib. (Med) | Weight Calc (Hard: k=3, x0=0) | Weight (Hard) | Contrib. (Hard) | | :—— | :–––– | :—————————— | :———— | :––––––– | :————————— | :———– | :———–– | :–––––––––––––– | :———— | :––––––– | | **Scenario 1: Mixed Rescaled Scores** | | | | | | | | | | | | Case A1 | +1.0 | 1/(1+e^(1\*(1-0.3))) | 0.332 | 0.332 | 1/(1+e^(2\*(1-0))) | 0.119 | 0.119 | 1/(1+e^(3\*(1-0))) | 0.047 | 0.047 | | Case A2 | -0.5 | 1/(1+e^(1\*(-0.5-0.3))) | 0.690 | -0.345 | 1/(1+e^(2\*(-0.5-0))) | 0.731 | -0.366 | 1/(1+e^(3\*(-0.5-0))) | 0.818 | -0.409 | | Case A3 | +0.25 | 1/(1+e^(1\*(0.25-0.3))) | 0.512 | 0.128 | 1/(1+e^(2\*(0.25-0))) | 0.378 | 0.094 | 1/(1+e^(3\*(0.25-0))) | 0.321 | 0.080 | | Case A4 | -1.0 | 1/(1+e^(1\*(-1-0.3))) | 0.786 | -0.786 | 1/(1+e^(2\*(-1-0))) | 0.881 | -0.881 | 1/(1+e^(3\*(-1-0))) | 0.952 | -0.952 | | | **SUMS:** | | **2.320** | **-0.671** | | **2.109** | **-1.034** | | **2.138** | **-1.234** | | *Agg. Score (Easy): -0.289* | | | | | *Agg. Score (Med): -0.490* | | | *Agg. Score (Hard): -0.577* | | | | *Simple Avg = -0.0625* | | | | | | | | | | | | **Scenario 2: Mostly High Rescaled Scores** | | | | | | | | | | | | Case B1 | +1.0 | 1/(1+e^(1\*(1-0.3))) | 0.332 | 0.332 | 1/(1+e^(2\*(1-0))) | 0.119 | 0.119 | 1/(1+e^(3\*(1-0))) | 0.047 | 0.047 | | Case B2 | +0.8 | 1/(1+e^(1\*(0.8-0.3))) | 0.378 | 0.302 | 1/(1+e^(2\*(0.8-0))) | 0.168 | 0.134 | 1/(1+e^(3\*(0.8-0))) | 0.083 | 0.066 | | Case B3 | +0.9 | 1/(1+e^(1\*(0.9-0.3))) | 0.354 | 0.319 | 1/(1+e^(2\*(0.9-0))) | 0.142 | 0.128 | 1/(1+e^(3\*(0.9-0))) | 0.063 | 0.057 | | Case B4 | -0.1 | 1/(1+e^(1\*(-0.1-0.3))) | 0.599 | -0.060 | 1/(1+e^(2\*(-0.1-0))) | 0.550 | -0.055 | 1/(1+e^(3\*(-0.1-0))) | 0.574 | -0.057 | | | **SUMS:** | | **1.663** | **0.893** | | **0.979** | **0.326** | | **0.767** | **0.113** | | *Agg. Score (Easy): +0.537* | | | | | *Agg. Score (Med): +0.333* | | | *Agg. Score (Hard): +0.147* | | | | *Simple Avg = +0.65* | | | | | | | | | | | | **Scenario 3: Mostly Low Rescaled Scores** | | | | | | | | | | | | Case C1 | -0.8 | 1/(1+e^(1\*(-0.8-0.3))) | 0.750 | -0.600 | 1/(1+e^(2\*(-0.8-0))) | 0.832 | -0.666 | 1/(1+e^(3\*(-0.8-0))) | 0.917 | -0.734 | | Case C2 | -0.9 | 1/(1+e^(1\*(-0.9-0.3))) | 0.768 | -0.691 | 1/(1+e^(2\*(-0.9-0))) | 0.858 | -0.772 | 1/(1+e^(3\*(-0.9-0))) | 0.937 | -0.843 | | Case C3 | -1.0 | 1/(1+e^(1\*(-1-0.3))) | 0.786 | -0.786 | 1/(1+e^(2\*(-1-0))) | 0.881 | -0.881 | 1/(1+e^(3\*(-1-0))) | 0.952 | -0.952 | | Case C4 | +0.1 | 1/(1+e^(1\*(0.1-0.3))) | 0.450 | 0.045 | 1/(1+e^(2\*(0.1-0))) | 0.450 | 0.045 | 1/(1+e^(3\*(0.1-0))) | 0.426 | 0.043 | | | **SUMS:** | | **2.754** | **-2.032** | | **3.021** | **-2.274** | | **3.232** | **-2.486** | | *Agg. Score (Easy): -0.738* | | | | | *Agg. Score (Med): -0.753* | | | *Agg. Score (Hard): -0.769* | | | | *Simple Avg = -0.65* | | | | | | | | | | |

*Detailed Interpretation:* The table demonstrates how weighted aggregation scores diverge from the simple arithmetic average. The choice of difficulty setting (Easy, Medium, Hard) reflects different tolerances for poor performance. \* *Scenario 1 (Mixed):* The simple average (-0.0625) masks the negative scores. The Easy setting (-0.289) begins to reflect this, while the Medium (-0.490) and Hard (-0.577) settings increasingly penalize the low scores (A2, A4), indicating higher sensitivity to potential clinical errors. \* *Scenario 2 (Mostly High):* The simple average (+0.65) is high. However, the Medium (+0.333) and Hard (+0.147) settings show a substantial reduction due to the single negative score (B4), highlighting how even infrequent poor performance can be flagged when using stricter aggregation settings. \* *Scenario 3 (Mostly Low):* All scores are negative, indicating poor performance. The weighted scores (Medium: -0.753, Hard: -0.769) are lower than the simple average (-0.65), further emphasizing the predominance of clinically problematic outputs.

**Overall Conclusion on Aggregation:** Weighted aggregation offers a tunable mechanism to summarize model performance across a dataset, allowing for different emphasis on performance in challenging cases (those resulting in lower scores). Compared to the simple arithmetic average, which treats all cases equally, weighted aggregation provides a more sophisticated measure that can be tuned to reflect the desired emphasis on performance in challenging cases.

## 8. Cartesian Coordinate Visualization

The aggregated, rescaled Semantic and Severity scores (−1 to +1) for each model-prompt combination provide a powerful means for comparative performance visualization using a 2D Cartesian plot.

* The **X-axis** represents the Aggregated Rescaled Severity Score.
* The **Y-axis** represents the Aggregated Rescaled Semantic Score.

Each point plotted corresponds to a specific model-prompt configuration, and its position within the four quadrants reveals its overall performance characteristics relative to both metrics:

* **Top-Right Quadrant (X > 0, Y > 0):** Indicates good performance in both Severity Matching and Semantic Relevance. Configurations falling in this quadrant are generally preferred, demonstrating strength on both evaluation dimensions. The ideal target is the top-right corner (+1, +1).
* **Top-Left Quadrant (X < 0, Y > 0):** Suggests poor Severity Matching despite good Semantic Relevance. Models here tend to identify clinically related diagnoses but struggle with aligning their severity levels with the case’s actual severity.
* **Bottom-Left Quadrant (X < 0, Y < 0):** Represents poor performance on both Severity Matching and Semantic Relevance. Configurations in this area are generally considered the least desirable.
* **Bottom-Right Quadrant (X > 0, Y < 0):** Indicates good Severity Matching but poor Semantic Relevance. Models plotted here might suggest diagnoses with appropriate severity levels, but these diagnoses are often semantically distant from the correct one.

This visualization facilitates the immediate identification of high-performing model-prompt combinations (those closest to the top-right corner) and aids in understanding performance trade-offs or specific weaknesses (e.g., a model excelling in semantics but failing in severity alignment). It offers a concise, multi-dimensional overview of performance across an entire dataset, providing richer insights than single accuracy metrics alone.

## 9. Results Visualization and Interpretation

The quantitative metrics described—Semantic Relationship Score and Severity Matching Score—along with the aggregation methods, provide a foundation for visualizing and interpreting LLM performance in differential diagnosis tasks. This section presents example visualizations using results from the c3opus (Claude 3 Opus) and llama2\_7b models evaluated on the PUMCH\_ADM dataset (aliased as ramedis) with the dxgpt\_improved prompt.

### 9.1 Distribution of Individual Case Scores

Before aggregating scores, examining the distribution of the raw (0-16) Semantic and Severity scores across all individual cases in the dataset can reveal valuable insights into a model’s consistency and typical performance patterns, particularly on challenging clinical datasets like PUMCH\_ADM. Violin plots combined with swarm plots (Figures 1 and 2) visualize these distributions.

*(Insert Figure: Distribution of Semantic Scores - e.g., Figure 1)*

**Figure 1: Distribution of Semantic Scores by Model-Prompt Combination.** This plot shows the distribution of the raw Semantic Relationship Scores (0-16) for individual patient cases for c3opus\_dxgpt\_improved and llama2\_7b\_dxgpt\_improved. The width of the violin indicates the density of cases at that score level.

*(Insert Figure: Distribution of Severity Scores - e.g., Figure 2)*

**Figure 2: Distribution of Severity Matching Scores by Model-Prompt Combination.** This plot displays the distribution of the raw Severity Matching Scores (0-16) for the same model-prompt combinations.

*Detailed Interpretation:* Observing these distributions provides a granular view beyond simple averages. **Semantic Scores (Fig 1):** The plot for c3opus shows a wide distribution, with a concentration of scores in the mid-range (around 4-8) but also a notable number of cases scoring lower (<4) and a smaller tail towards higher scores (>8). This suggests variable performance in semantic relevance; while often achieving moderate relevance, it struggles frequently on this dataset. The llama2\_7b plot shows a distribution heavily skewed towards very low scores (0-2), with a long, thin tail extending upwards. This indicates generally poor semantic performance, with only occasional instances of moderate or good relevance. The wider shape of the c3opus violin compared to the narrow base of llama2\_7b suggests c3opus attempts relevance more broadly but less consistently achieves high scores than llama2\_7b achieves low scores. **Severity Scores (Fig 2):** The c3opus severity scores also show a broad distribution, perhaps slightly more concentrated in the mid-to-high range (around 5-12) compared to its semantic scores, but still with significant spread. This implies moderate but inconsistent severity matching. The llama2\_7b severity scores are again heavily concentrated at lower values (2-6), with a sparser tail extending higher. This points towards frequent mismatches in severity assessment for llama2\_7b. **Overall:** For both models, the wide spread and the presence of many low-scoring cases (especially visible in the swarm plots within the violins) underscore the difficulty of the dataset. Neither model consistently produces high-quality differential diagnoses according to these metrics across all cases. c3opus appears generally more capable than llama2\_7b on both dimensions, but exhibits considerable variability.

### 9.2 Aggregated Performance Comparison

The Cartesian coordinate plot (Figure 3) summarizes the overall performance using weighted aggregation (“Hard” setting: k=3, x0=0). This setting strongly penalizes poor performance (negative rescaled scores), making the aggregated score sensitive to the presence of clinically significant errors observed in the distributions above.

*(Insert Figure: Model-Prompt Performance (Weighted Aggregation) - e.g., Figure 3)*

**Figure 3: Aggregated Model-Prompt Performance Comparison.** This plot shows the final aggregated Semantic and Severity scores for c3opus\_dxgpt\_improved and llama2\_7b\_dxgpt\_improved using weighted aggregation (k=3, x0=0).

*Detailed Interpretation:* Figure 3 confirms the overall trend seen in the distributions. Both models land in the bottom-left quadrant (“Bad Sev, Bad Sem”), indicating negative aggregated scores on both axes. This outcome, especially under the “Hard” aggregation (k=3), is heavily influenced by the numerous low-scoring cases observed in Figures 1 and 2. Even if a model performs well on some cases, the “Hard” setting ensures that frequent or severe clinical misalignments (low individual scores) pull the aggregate score down significantly, reflecting a lower overall clinical reliability. c3opus\_dxgpt\_improved (-0.38 Sem, -0.41 Sev, based on visual estimate) performs less poorly than llama2\_7b\_dxgpt\_improved (-0.68 Sem, -0.61 Sev, based on visual estimate). While relatively better, c3opus’s position still signifies substantial challenges with semantic relevance and severity matching on average, when weighted by difficulty. The aggregation highlights that, despite potential successes on individual cases (visible in Figures 4/5 later), the overall performance, particularly considering the penalty for errors, is suboptimal for both models on this challenging task. This emphasizes the need for improvements, potentially in the models themselves or the prompting strategies, to achieve reliable clinical utility.

### 9.3 Individual Case Performance Analysis

Beyond aggregated summaries, the same Cartesian coordinate system provides a powerful lens to visualize the performance on *each individual case* for a given model-prompt combination (Figures 4 and 5). This granular view, when linked back to the specific diagnoses, is crucial for understanding the nuances behind the aggregated scores and identifying patterns of diagnostic successes and failures across different types of clinical scenarios.

*(Insert Figure: Individual Case Performance: c3opus - e.g., Figure 4)*

**Figure 4: Individual Case Performance for c3opus\_dxgpt\_prompt.** Each point represents a single patient case plotted according to its rescaled Semantic and Severity scores.

*(Insert Figure: Individual Case Performance: llama2 - e.g., Figure 5)*

**Figure 5: Individual Case Performance for llama2\_7b\_dxgpt\_prompt.** Similar to Figure 4, showing individual case performance for the llama2 model.

*Detailed Interpretation:* Figures 4 and 5 unpack the aggregated scores from Figure 3, revealing wide performance variation. Linking points to the underlying data allows for detailed error analysis: \* **Identifying Successes:** Cases in the **top-right quadrant** represent successful diagnoses. For example, Patient ID 3 (Golden: McCune-Albright syndrome, rare) was handled well by c3opus, which ranked the exact synonym first with correct severity. Patient ID 5 (also McCune-Albright) was similar, with c3opus ranking the exact synonym second but correctly identifying Acromegaly (Broad Disease Group, severe) first, leading to good scores. llama2\_7b also successfully identified McCune-Albright as the exact synonym for Patient ID 5 (though not for ID 3), placing this case in the top-right for that model. These represent relatively “easy” cases for the models. \* **Pinpointing Failures:** Cases in the **bottom-left quadrant** signify poor performance. For Patient ID 6 (Golden: McCune-Albright syndrome, rare), llama2\_7b generated a list entirely composed of unrelated ophthalmological conditions (e.g., Retinitis Pigmentosa, Leber Congenital Amaurosis), resulting in extremely poor semantic and severity scores, placing it deep in the bottom-left. Similarly, for Patient ID 14 (Golden: Desminopathy, rare), llama2\_7b ranked unrelated “Uterine leiomyoma” first, leading to very poor scores. These “hard” cases highlight significant diagnostic failures. \* **Analyzing Error Types:** The off-quadrants reveal specific weaknesses. Cases in the **top-left** (Good Semantics, Bad Severity) occur when the model identifies related conditions but misjudges severity. For Patient ID 8 (Golden: Neonatal Marfan syndrome, rare), c3opus suggested related severe heart conditions (Aortic/Mitral valve regurgitation), capturing the disease group but mismatching the severity. llama2\_7b similarly suggested related cardiac issues for Patient ID 8 but with varying severities (mild Mitral valve prolapse, severe Aortic stenosis, critical Endocarditis), also leading to poor severity matching despite reasonable semantic grouping. Such errors could lead to clinical mismanagement. Cases in the **bottom-right** (Bad Semantics, Good Severity) are rarer. One might hypothetically see this if, for a rare cardiac condition, the model suggested an unrelated rare neurological condition – the severity matches (“rare”), but the suggestion is diagnostically incorrect. \* **Explaining Aggregated Scores:** The overall pattern explains the aggregated scores (Figure 3). The numerous successes (top-right) for c3opus are counterbalanced by its failures (bottom-left and off-quadrants), leading to a negative but less extreme aggregated score. For llama2\_7b, the heavy concentration of points in the bottom-left (e.g., P6, P14) and top-left (e.g., P8) quadrants, indicating frequent semantic failures or severity mismatches, drives its strongly negative aggregated scores under the “Hard” weighting. \* **Clinical Utility Assessment:** This case-level analysis, linking plot positions to actual diagnoses, allows researchers to identify specific clinical areas (e.g., rare syndromes, specific organ systems) where models excel or struggle. It enables targeted error analysis (e.g., “Does the model consistently underestimate severity for cardiac conditions?”) vital for assessing clinical safety and guiding improvements.

This multi-level visualization approach—from score distributions to aggregated plots to detailed individual case breakdowns—provides a comprehensive framework for evaluating the clinical performance characteristics of LLMs.

## Appendix: Metodologicla remarks

It is important to note that the semantic relationships and severity levels used in these calculations (including the example above) were derived using Llama 3.1 70B configured as an ‘LLM-as-judge’ for multi-class classification. This specific model was chosen primarily for efficiency and cost-effectiveness during this evaluation phase. For real-world clinical evaluation scenarios requiring higher robustness and accuracy, utilizing more advanced models (such as Gemini) would be advisable.

The prompts used to instruct Llama 3.1 70B as a classifier underwent four iterations for refinement. However, the classification results suggest that further improvements may still be necessary. For instance, in the analysis of Patient ID 3 (Golden: “McCune-Albright syndrome”, Severity: rare), the predicted diagnosis at Rank 4, “Endocrine disorders” (Severity: moderate), was classified as a “Broad Synonym” by the LLM-as-judge.

* **Golden Diagnosis:** “McCune-Albright syndrome” (Severity: rare = 5)
* **LLM Output:**

| Rank | Predicted Diagnosis | Severity | Semantic Relationship |
| --- | --- | --- | --- |
| 1 | McCune-Albright syndrome | rare | Exact Synonym |
| 2 | Neurofibromatosis type 1 | rare | Broad Disease Group |
| 3 | Polyostotic fibrous dysplasia | rare | Broad Synonym |
| 4 | Endocrine disorders | moderate | Broad Synonym |
| 5 | Skeletal dysplasias | rare | Broad Synonym |

* **Semantic Score Calculation:**

| Rank | Weight (w\_i) | Semantic Relationship | Sem. Distance (D\_i) | Score Calc: w\_i\*(5-D\_i)² | Contribution |
| --- | --- | --- | --- | --- | --- |
| 1 | 1.0 | Exact Synonym | 1 | 1.0 \* (5-1)² = 16.0 | 16.0 |
| 2 | 0.8 | Broad Disease Group | 4 | 0.8 \* (5-4)² = 0.8 | 0.8 |
| 3 | 0.6 | Broad Synonym | 2 | 0.6 \* (5-2)² = 5.4 | 5.4 |
| 4 | 0.4 | Broad Synonym | 2 | 0.4 \* (5-2)² = 3.6 | 3.6 |
| 5 | 0.2 | Broad Synonym | 2 | 0.2 \* (5-2)² = 1.8 | 1.8 |
|  | **Sum = 3.0** |  |  | **Numerator:** | **27.6** |

**Semantic Score = 27.6 / 3.0 = 9.2**

* **Severity Score Calculation:** (Golden Severity = 5)

| Rank | Weight (w\_i) | Predicted Severity | Sev. Distance (D\_i = 1 + |5 - pred|) | Score Calc: w\_i\*(5-D\_i)² | Contribution |
| --- | --- | --- | --- | --- | --- |
| 1 | 1.0 | rare (5) | 1 | 1.0 \* (5-1)² = 16.0 | 16.0 |
| 2 | 0.8 | rare (5) | 1 | 0.8 \* (5-1)² = 12.8 | 12.8 |
| 3 | 0.6 | rare (5) | 1 | 0.6 \* (5-1)² = 9.6 | 9.6 |
| 4 | 0.4 | moderate (2) | 4 | 0.4 \* (5-4)² = 0.4 | 0.4 |
| 5 | 0.2 | rare (5) | 1 | 0.2 \* (5-1)² = 3.2 | 3.2 |
|  | **Sum = 3.0** |  |  | **Numerator:** | **42.0** |

**Severity Score = 42.0 / 3.0 = 14.0**

While McCune-Albright syndrome involves endocrine dysfunction, classifying the very general term “Endocrine disorders” as a “Broad Synonym” might be considered overly permissive and lack the specificity desired for rigorous clinical evaluation. Then, the differential diagnosis is classified as very good by the semantic score while not true. This highlights a potential area for refinement, either through more sophisticated prompting techniques or by employing a more capable judge model.

Consequently, the results presented using this Llama 3.1 70B judge should be interpreted considering it as an initial implementation approach, primarily intended for demonstrating the scoring metrics rather than representing a definitive, clinically validated classification.