Task-1 Guidelines

▼ **Instruction**

You are an expert meta-reviewer. **Evaluate the following peer review text according to the scoring methodology.**

Evaluation Process:

1. **Read the given peer review** generated for a paper submitted to the ICLR conference.

2. **Give rating for the categories described below** on the full scale from **1 (lowest rating)** to **5 (highest rating)**.

3. **Assign scores independently per category** (they may differ significantly).

Category 1: Reviewer Confidence
1 – Unqualified: Admits lack of expertise │ Fundamental misunderstandings │ Fails to engage with technical content
2 – Limited: Surface-level comprehension │ Misses >2 key concepts │ Confuses established methods
3 – Functional: Grasps main contributions │ Identifies 1-2 technical flaws │ Limited literature context
4 – Qualified: Precise technical critique │ Contextualizes in recent work (>3 citations) │
Identifies methodological tradeoffs
5 – Authoritative: Demonstrates field leadership │ Predicts follow-up research directions │
Reveals fundamental limitations in approach

Category 2: Review Thoroughness
1 – Negligent: <3 substantive comments │ Skips methodology/results │ Fails to address claims
2 – Patchy: Only addresses obvious aspects │ 0 analysis of experiments │ No ethical consideration
3 – Standard: Checks main claims │ Basic methodology review │ 1

reproducibility concern noted
4 – Rigorous: Line-by-line methodology check │ Statistical validity analysis │ 2+ improvement suggestions
5 – Forensic: Alternative approach proposals │ Re-analysis of key results │ Compliance checklist (ethics, reproducibility, etc.)

Category 3: Constructiveness
1 – Destructive: Personal attacks │ Unsubstantiated dismissal │ 0 actionable feedback
2 – Demanding: Unrealistic requests ("redo all experiments") │ Vague quality complaints
3 – Directive: General improvement areas │ 1-2 specific examples │ Mixed positive/negative tone
4 – Strategic: Priority-ranked suggestions │ Alternative methodology paths │ Conference-specific improvements
5 – Transformative: Step-by-step revision roadmap │ Code/experiment snippets provided │ Before-after examples

Category 4: Helpfulness
1 – Misleading: Factually incorrect suggestions │ Contradicts paper's goals │ Harmful recommendations
2 – Abstract: "Improve writing" without examples │ "Add experiments" without design
3 – Guided: Section-specific edits │ 2-3 implementable suggestions │ Partial error diagnosis
4 – Impactful: Directly address weaknesses │ Templates/examples provided │ Clear priority levels
5 – Revelatory: Exposes fundamental flaws │ Provides benchmark comparisons │ Supplies missing literature

▼ **Scoring Methodology**

For EACH category, assign scores based on,

1. Find the HIGHEST level (5-1) where ALL criteria are FULLY satisfied (Base Level)

2. Check criteria in the NEXT HIGHER LEVEL (Base Level + 1):

   - Calculate partial credit: (Satisfied Criteria in Higher Level / Total Criteria in Higher Level)

3. Final Score = Base Level + Partial Credit

4. Round to 2 decimal places (0.33 increments for 3-criteria levels, 0.5 for 2-criteria levels)

▼ **Expected Output**

"review_scores": {{
"Category 1": "",
"Category 2": "",
"Category 3": "",
"Category 4": "" }}

▼ Example

When you are reading the review, say for **Category-1 (Reviewer Confidence)**

- The **HIGHEST level (5-1)** where **ALL criteria are FULLY satisfied** is **level 3 (Base Level)**.

  This means all the "Functional" criteria are present in the given review namely,

   - Grasps main contributions

   - Identifies 1-2 technical flaws

   - Limited literature context

  ⇒ **Base Level = 3**

- Check criteria in the **NEXT HIGHER LEVEL (Base Level + 1)**
  But, all the criteria to assign a score of 4 ("Qualified" criteria) are not present in the given review

   - Precise technical critique - PRESENT✅

   - Contextualizes in recent work (>3 citations) - NOT PRESENT ❌

   - Identifies methodological tradeoffs - PRESENT✅

     Hence, as only 2 out of 3 criteria are present

⇒ **Partial score = 0.33∗2** (for this level)

- **Final score** : 3 (Base Level) + 0.33∗2 (Partial Credit) = **3.67**

Hence,

Category 1 : 3.67

Similarly scores are calculated for the remaining 3 categories.

Note that:

- 0.5 is the partial credit per criteria for levels having only two criteria.