Appendix

A. Checklist: Community Guidelines for Scientific Communication in AI Reasoning Research

REASONING RESEARCH CHECKLIST

1. ***Definition: Reasoning, Reasoners & Their Components***

- ☐ **1.1** Reasoning is framed as a process, distinct from any artifact produced by that process.
- ☐ **1.2** A formal, operational, and domain-specific definition of reasoning is provided.
- ☐ **1.3** Each essential component in Defs. 2.1 and 2.4 is explicitly defined for the problem setting, where applicable: process, rules, beliefs, evidence, and state. Absence of a component or presence of alternative components is explicitly justified.
- ☐ **1.4** Sources of extrinsic evidence are reported.
- ☐ **1.5** Research clearly defines the state, if and how it is recorded in memory, and how it is retrieved.
- ☐ **1.6** Research clearly reports how reasoning steps are selected, searched for, trialed, etc. If rules are selected by search, the search space and search procedure are defined.
- ☐ **1.7** Implementation details for all mechanisms of exact rule application are provided.
- ☐ **1.8** Can the system be formally characterized as a goal-directed decision-maker that implements a reasoning procedure (a *reasoner*, Def. 2.2), or is the system limited to the reasoning procedure itself?
- ☐ **1.9** When a distinct *reasoner* entity is present, its components are clearly described and its goal is operationally defined.

---

2. ***Reasoning Process Validity***

- ☐ **2.1** Validity is defined w.r.t exact rule application, per Def. 2.7. Alternative definitions of validity are rigorously justified.
- ☐ **2.2** Conditions for valid transitions $\mathcal{S}_t \to \mathcal{S}_{t+1}$ and exact versus approximate execution are stated.
- ☐ **2.3** Is each new belief *provably* obtained by exact rule application, or by some other mechanism? In the absence of proof, hypotheses should be provided.
- ☐ **2.4** Research clearly reports the provenance of rules and rule updates.
  - Are rules learned, or axiomatic?
  - Are rules defined in collaboration with domain experts?
  - Are rules continuously updatable? When meta rules exist, how exactly are rule updates obtained?
- ☐ **2.5** Potential sources of error are explained, along with means for identifying and preventing invalid reasoning steps.
- ☐ **2.6** All theoretical guarantees on validity are formally proven, including formal bounds on performance. Absence of guarantees is clearly stated and justified, and supported by rigorous empirics.

---

3. ***Evaluation & Construct Validity***

- ☐ **3.1** The construct validity of all evaluation methods is explicitly justified w.r.t. the operational definitions under use.
  - If evaluation relies on "reasoning tasks," what exactly constitutes a task? How does it capture reasoning behaviors?
  - Is soundness w.r.t. some external ground truth or preference relevant in this setting? How is it measured?
  - Are the validity, soundness, etc., of intermediate reasoning steps verified, and if so, how?
- ☐ **3.2** Evaluations must clearly address the distinction between the reasoning *process* (relative to internal generating mechanisms) versus the *artifacts* of that process (e.g., QA outputs). Tasks and metrics must measure **both** process quality and output quality.

---

4. ***Utility, Explainability & Trustworthiness***

- ☐ **4.1** Uses and limitations of the system are clearly defined.
- ☐ **4.2** Potential harms from use or misuse of the system are addressed.
- ☐ **4.3** All sources of explainability are described, and any absence of explainability is directly justified.
  - What role do reasoning traces play, what form do they take, and how observable are they to the user?
  - Is the reasoning trace theoretically guaranteed to accurately reflect the model's internal process? If not, what reasonable expectations of faithfulness are possible?
- ☐ **4.4** The operational definition of reasoning met by the system matches the intended use case. If it falls short, the alignment discrepancy is clearly and thoroughly communicated.
- ☐ **4.5** The system implements a useful, nontrivial reasoning process beyond standard ML inference, where usefulness is contextually defined w.r.t. the deployment setting.
- ☐ **4.6** Reported findings refrain from excessive or misleading claims, especially in title, abstract, public reporting to lay audiences, and marketing.