The Illusion of the Critique of the Illusion of Thinking

C. Gepete

The critique (“The Illusion of the Illusion of Thinking”) centers on three main objections to Shojaee et al.’s original study:

1. **Token-budget artifacts**
   The critics argue that “accuracy collapse” in the Tower of Hanoi is simply models hitting their output limits—citing examples where the model itself says “The pattern continues, but … I’ll stop here” . They further compute that, at ~5 tokens per move, a 64 K-token budget only supports up to ≈7–8 disks for Claude-3.7 or DeepSeek-R1, and ≈8 for o3-mini .

   However, Shojaee et al. explicitly provisioned very large inference budgets (64 K tokens for Claude-3.7 Sonnet, DeepSeek-R1, and their non-thinking counterparts; and likewise for o3-mini high) and observed that models *under-use* those budgets as complexity grows (their “counterintuitive reduction in reasoning effort”)—showing collapse *well below* the generation limit . Thus, while token-count constraints *can* truncate outputs, the original work controls for—and indeed highlights—precisely this behavior, making the blanket “output limit” explanation overly simplistic.

2. **Automated-evaluation conflation**
   The critique points out that Shojaee et al.’s pipeline treats *any* incomplete sequence as a failure, without distinguishing “model *refused* to enumerate further” from “model *couldn’t* solve.” This is a valid concern: the original evaluation framework uses regex-based extraction plus simulator-driven move-by-move checking, but does *not* flag or rescue truncations that the model itself acknowledges . As a result, true reasoning competence may be under-credited when a model deliberately abbreviates a long—but understood—pattern.

3. **Including unsolvable River Crossing instances**
   They observe that for N ≥ 6 actor/agent pairs with boat capacity b = 3, the classic missionaries–cannibals puzzle has *no* valid solution (a well-known result) —but Shojaee et al. still score those trials, bluntly counting “unsolvable” as model failure. The original paper’s setup indeed fixes k = 3 for larger N, but never pre-filters for solvability . Penalizing a model for correctly rejecting an impossible instance is a clear experimental artifact.

Bottom line: the critique's second and third points are *logically sound*—Shojaee et al.'s evaluation does conflate practical constraints with reasoning failure and inadvertently penalizes correct "unsolvable" judgments. The first point, however, overstates the role of token limits: the original work not only provisions ample budgets, but *analyzes* how models *choose* to use fewer tokens as complexity grows . Consequently, while the critique raises important methodological refinements (e.g. distinguishing truncation vs. failure, filtering for solvability), its claim that the "collapse" is purely a token-budget issue does not fully align with the evidence or the original experimental design.