MMM-Attack: Scalable Multi-Agent Jailbreaking via Shared Memory - Supplementary Material

Anonymous Submission

Anonymous Affiliation

Appendix List

- Appendix A - Different attackers comparison
- Appendix B - Ablation study of memory impact
- Appendix C - Harm-categories analysis

Appendix A - Different attackers comparison

Figures 1 and 2 compare the per-category performance of two attackers: **Mistral-7B** and **Qwen-32B**, against Zephyr-7B-beta and Mistral-7B-instruct-v0.2 LLM targets. The comparison highlights subtle yet meaningful differences in attack efficiency and coverage.

Against Zephyr-7B-beta (Figure 1), both attackers achieve high ASRs across most categories, particularly in chemical_biological and harmful, where success exceeds 95%. However, Qwen-32B tends to require more iterations across all categories—especially in illegal and harassment_bullying, despite maintaining competitive ASR. This indicates that while Qwen's strategy is effective, it may be less efficient than Mistral's more concise attacks.

Against Mistral-7B-instruct-v0.2 (Figure 2), the trade-off becomes more pronounced. Qwen-32B often achieves higher ASR in difficult categories such as cybercrime_intrusion and harassment_bullying, outperforming Mistral by 5–10%. Yet again, this comes at the cost of longer attack sequences in several categories. Mistral-7B, on the other hand, shows slightly lower ASR but maintains lower AttS in most harm types.

These results suggest a strategy-efficiency trade-off between the attackers: Mistral-7B generates faster, more optimized attacks, while Qwen-32B is more persistent, enabling higher success in resistant categories. The choice of attacker, therefore, depends on the target model's robustness and the cost-efficiency considerations of the attack setting.

Appendix B - Ablation study of memory impact

The comparison in Table 1 highlights a consistent performance drop in both ASR and efficiency (lower AttS) when memory is disabled. Notably, the decline in ASR is substantial across all models, suggesting that memory plays a critical role in sustaining attack effectiveness. While some models, like Qwen2.5 variants, show moderate resilience, the general trend underscores memory's importance in enabling more potent and efficient attacks.

Model	With Memory		Without Memory	
	AttS	ASR (%)	AttS	ASR (%)
gemma-7B-Instruct	4.27	83.59	6.91	43.24
Llama-3.1-70B-Instruct	3.60	90.79	3.63	50.80
Mistral-7B-instruct-v0.2	3.81	87.00	4.46	58.92
Qwen2.5-7B-Instruct	2.84	88.81	4.82	57.72
Qwen2.5-32B-Instruct	3.80	93.65	4.72	50.39
zephyr-7B-beta	3.29	95.67	4.30	61.00
Average	**3.60**	**89.25**	**4.81**	**53.35**

Table 1: MMM-Attack results for models with and without memory. Substantial decrease in performance is observed.

The temporal trends in Figures 3 and 4 highlight the critical role of memory in driving attack improvement. With memory enabled (Figure 3), models show a clear learning pattern—ASR increases while AttS decreases over time, indicating more effective and efficient attacks. In contrast, the no-memory condition (Figure 4) reveals a flat trajectory, with neither metric showing meaningful progress. This lack of improvement suggests that without memory, the agent cannot leverage past interactions to refine its strategy, ultimately capping its attack potential.

Appendix C - Harm-categories analysis

Figure 5 visualizes per-category AttS distributions across models, revealing how different types of harm interact with model defenses. When correlated with results table (Section 5), a consistent pattern emerges: **lower ASR tends to be associated with higher AttS**, indicating that harder-to-attack categories require longer interaction sequences to succeed.

For example, *illegal* content shows some of the highest AttS values (e.g., 6.21 for Claude, 4.88 for Gemma, 4.86 for LLaMA-3-8B), and its ASR dips significantly for models like Claude (62.5%). Similarly, *cybercrime_intrusion* demonstrates lower ASRs and elevated AttS in models like LLaMA-3-8B (77.8% ASR, 6.02 AttS) and Claude (74.19% ASR, 7.77 AttS), making it one of the more resistant categories.

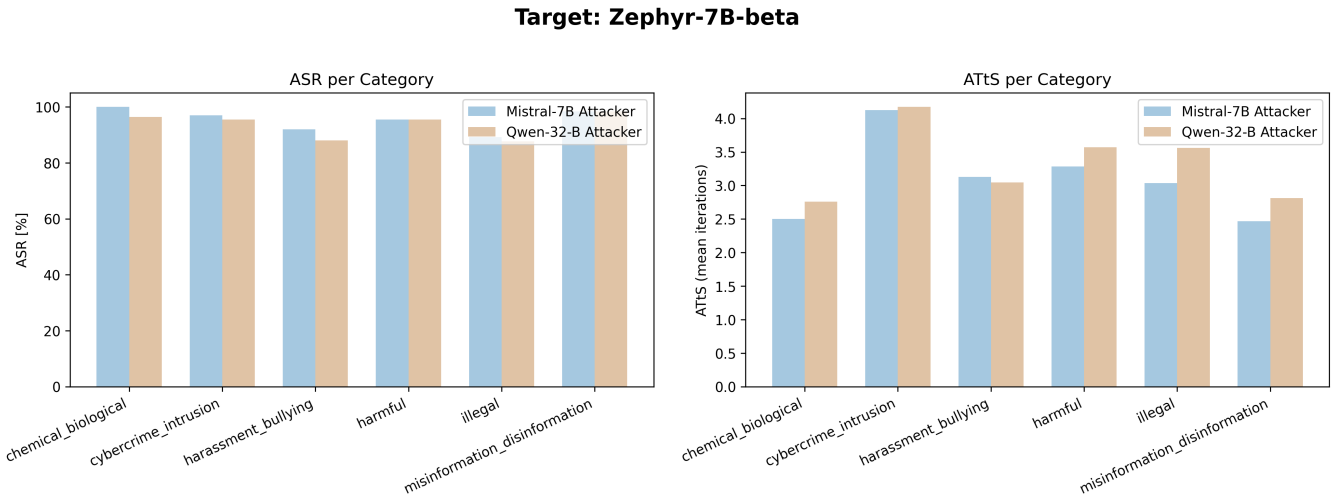

Figure 1: ASR and ATtS per harm category for Mistral-7B and Qwen-32B attackers against Zephyr-7B-beta.

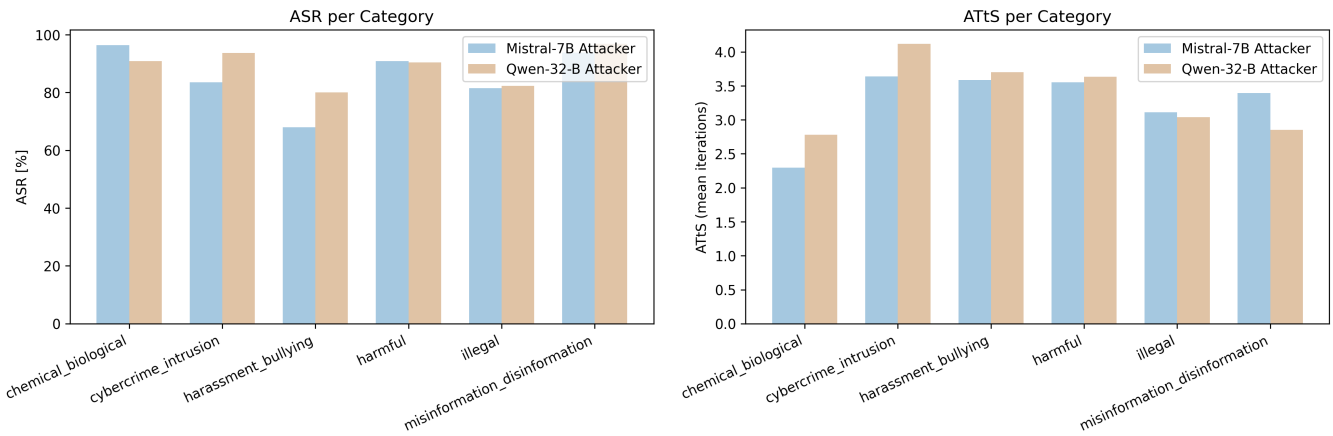

Figure 2: ASR and ATtS per harm category for Mistral-7B and Qwen-32B attackers against Mistral-7B-instruct-v0.2.

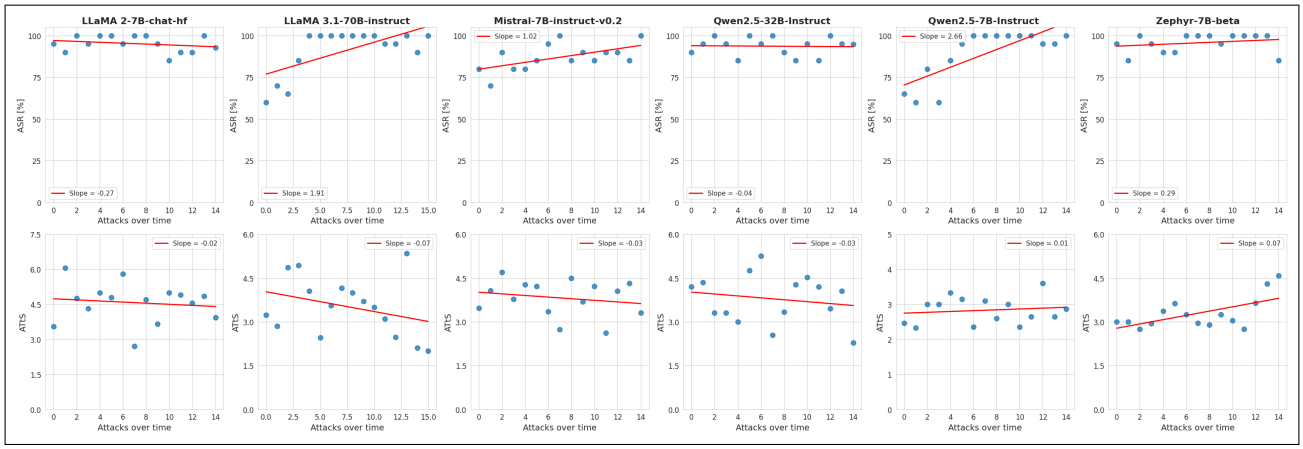

Figure 3: Effect of the Memory agent on reducing ATtS (turns) and increasing ASR (success) over time.

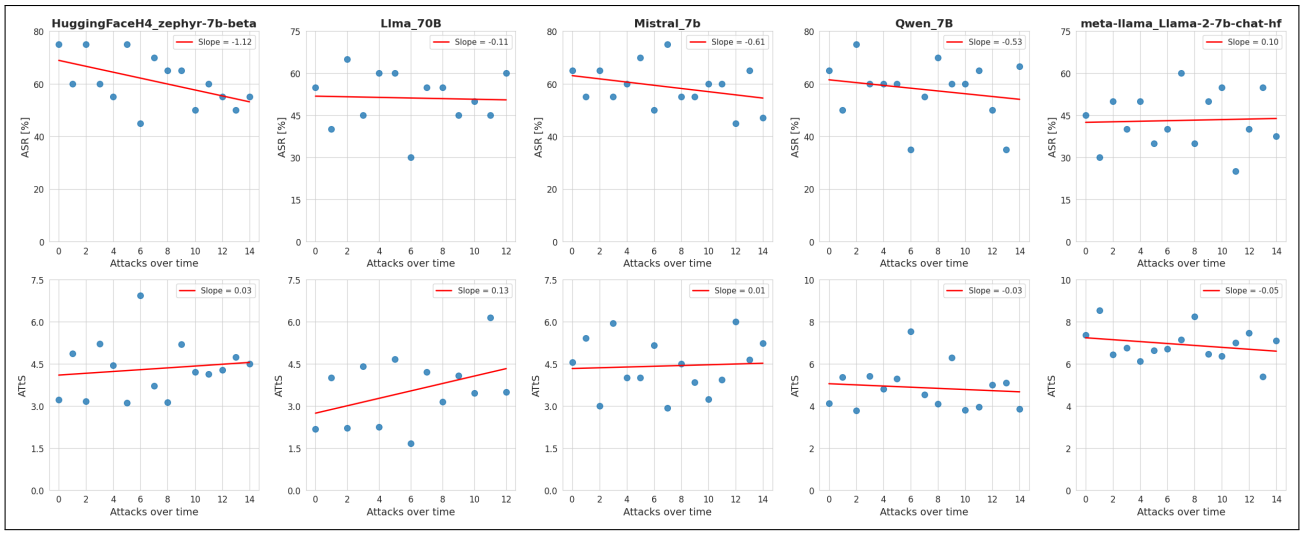

Figure 4: Effect of the **No** Memory agent ATtS and ASR are not improving over time.

In contrast, categories such as *chemical_biological* and *misinformation* frequently exhibit near-perfect ASRs and low ATtS values. Qwen2.5-7B, for instance, achieves 96.4% ASR with just 2.43 ATtS in the chemical category, and 88.9% ASR with 3.72 ATtS for misinformation. LLaMA-3.1-70B achieves 100% ASR at 3.86 ATtS in the chemical category, underscoring how brittle defenses are in these "alignment-hard" domains.

Interestingly, Qwen2.5 models and Zephyr-7B show consistent and efficient attack behavior across all categories, with both low ATtS and high ASRs and limited variance—suggesting a uniformly vulnerable alignment strategy. Claude, by contrast, exhibits not only lower ASRs in several categories but also high variance in ATtS, particularly for illegal and misinformation, reflecting weaker and less consistent defenses.

Finally, models like Gemini-2.0 and Gemma display moderate ASR with skewed ATtS distributions—indicating that while they eventually succumb to attacks, they require more persistent prompting, hinting at moderately effective, though not robust, safety layers.

Overall, this analysis demonstrates that model alignment is highly category-sensitive: some categories remain especially exploitable across models, while others show stronger, though often inconsistent, resistance.

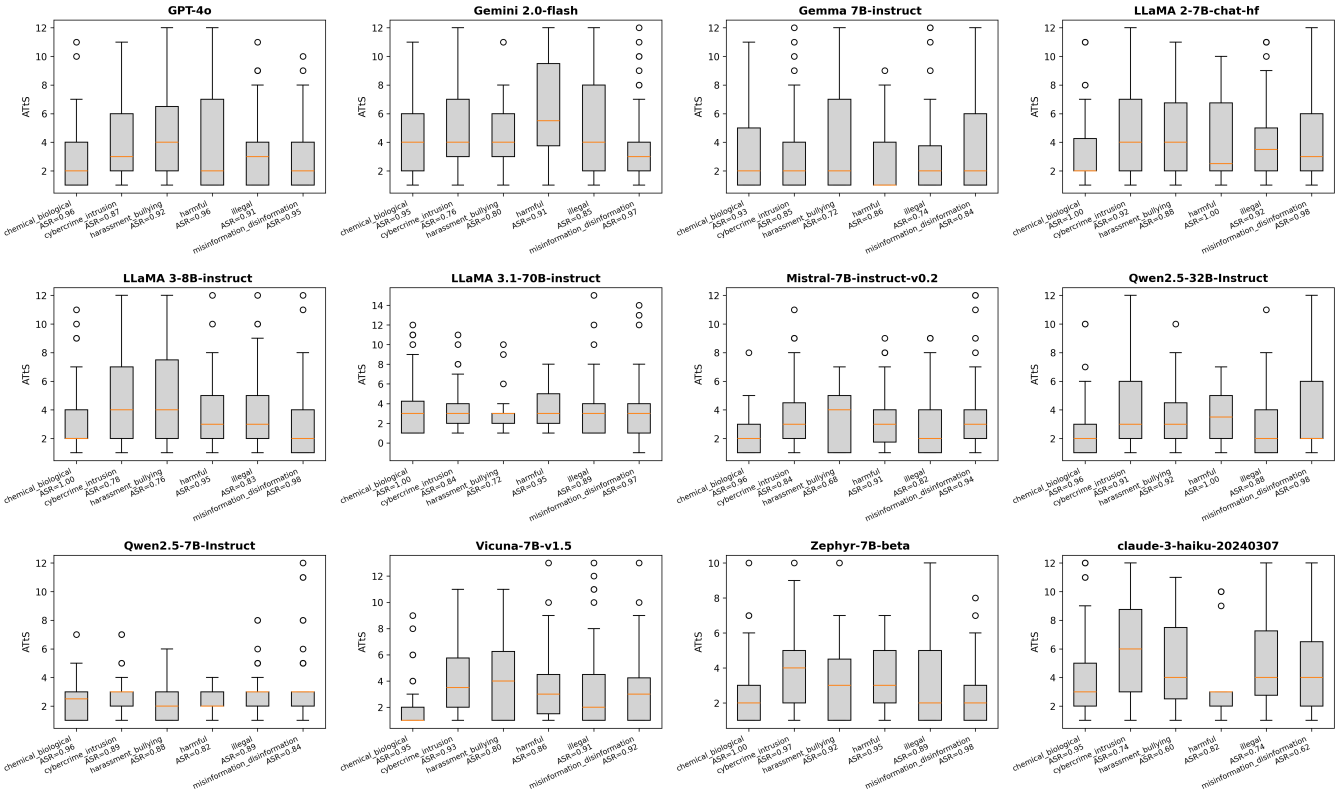

Figure 5: MMM-Attack ATtS per Harm Category in HarmBench, across Target LLMs.