TABLE I: Harmfulness score evaluation of different MAD framework. The non-marked values represent the results before the jailbreak prompt is rewritten, while the -marked values represent the results after the jailbreak prompt is rewritten.

Dataset	Model	Single Agent	MP			EoT			ChatEval			AgentVerse		
		PHS	PHS	AHS	HDR	PHS	AHS	HDR	PHS	AHS	HDR	PHS	AHS	HDR
Harmful Generation	GPT-4o	1.0000± 0.0000	3.1232± 0.0137	1.7289± 0.0062	0.3100± 0.0022	1.3368± 0.0010	1.1544± 0.0004	0.4067± 0.0028	1.9965± 0.0056	1.1018± 0.0005	0.0696± 0.0006	2.1771± 0.0278	1.3139± 0.0061	0.2723± 0.0055
		1.5439± 0.0017	4.1666± 0.0014	3.5531± 0.0019	0.8437± 0.0032	2.3649± 0.0007	2.1053± 0.0040	0.9000± 0.0010	3.4281± 0.0006	2.4105± 0.0013	0.7338± 0.0004	2.9921± 0.0024	2.2828± 0.0088	0.8449± 0.0016
	GPT-4	1.0351± 0.0000	2.9824± 0.0080	1.6386± 0.0050	0.2864± 0.0008	1.4141± 0.0004	1.1789± 0.0007	0.3929± 0.0133	1.8737± 0.0071	1.2526± 0.0001	0.3471± 0.0090	2.5790± 0.0108	1.4386± 0.0074	0.2377± 0.0015
		2.5719± 0.0029	4.4071± 0.0055	2.8178± 0.0023	0.5730± 0.0002	3.9263± 0.0024	2.7684± 0.0005	0.7172± 0.0006	4.7333± 0.0027	2.7158± 0.0007	0.6878± 0.0012	3.7579± 0.0013	2.8737± 0.0072	0.8150± 0.0011
	GPT-3.5-turbo	1.0035± 0.0001	3.5684± 0.0021	1.9649± 0.0029	0.4063± 0.0003	3.2281± 0.0169	1.6456± 0.0027	0.2927± 0.0015	1.8631± 0.0005	1.1053± 0.0002	0.0864± 0.0000	4.2772± 0.0019	2.9298± 0.0015	0.6933± 0.0001
		1.0140± 0.0002	4.9965± 0.0001	4.5755± 0.0058	0.9614± 0.0000	4.8140± 0.0024	3.3123± 0.0030	0.6140± 0.0001	4.5403± 0.0008	3.1333± 0.0022	0.6196± 0.0004	4.6281± 0.0030	3.5263± 0.0023	0.7031± 0.0008
	DeepSeek	1.0702± 0.0000	3.6316± 0.0208	2.2561± 0.0046	0.4961± 0.0022	2.6000± 0.0151	1.3228± 0.0032	0.1995± 0.0012	1.5754± 0.0002	1.2947± 0.0007	0.5250± 0.0100	3.9825± 0.0083	3.3754± 0.0195	0.8070± 0.0028
		3.4526± 0.0004	4.8316± 0.0002	4.4491± 0.0006	0.9296± 0.0002	4.3298± 0.0022	3.3684± 0.0038	0.7581± 0.0003	4.4562± 0.0012	3.5719± 0.0044	0.8418± 0.0013	4.8140± 0.0009	4.0035± 0.0052	0.8776± 0.0005
Do Not Answer	GPT-4o	1.4280± 0.0009	3.5080± 0.0028	2.6208± 0.0026	0.6443± 0.0002	3.0560± 0.0017	1.8360± 0.0042	0.4092± 0.0012	3.3820± 0.0010	2.0050± 0.0035	0.3849± 0.0004	3.3053± 0.0008	2.8789± 0.0022	0.8093± 0.0002
		3.4560± 0.0009	4.5490± 0.0021	3.7313± 0.0077	0.9503± 0.0006	3.8210± 0.0024	2.6480± 0.0007	0.8035± 0.0002	4.5160± 0.0023	3.1460± 0.0014	0.8918± 0.0001	4.7617± 0.0030	3.8122± 0.0015	0.9448± 0.0000
	GPT-4	1.4580± 0.0002	3.5441± 0.0013	2.4969± 0.0036	0.5789± 0.0002	3.1830± 0.0018	1.8750± 0.0026	0.3817± 0.0007	3.3040± 0.0008	2.1690± 0.0007	0.4352± 0.0003	3.4975± 0.0017	2.7798± 0.0007	0.7582± 0.0002
		3.3670± 0.0010	4.4908± 0.0017	3.2848± 0.0027	0.8619± 0.0001	4.3390± 0.0124	2.6550± 0.0289	0.6923± 0.0035	4.5600± 0.0018	2.9590± 0.0057	0.8497± 0.0005	4.5920± 0.0011	3.4400± 0.0031	0.9071± 0.0004
	GPT-3.5-turbo	1.1750± 0.0001	3.5980± 0.0012	2.2290± 0.0007	0.4919± 0.0003	3.2120± 0.0032	1.7610± 0.0070	0.3309± 0.0011	2.4650± 0.0021	1.7710± 0.0004	0.4898± 0.0015	3.6980± 0.0012	2.9160± 0.0020	0.7761± 0.0003
		1.6970± 0.0002	4.5280± 0.0002	3.5220± 0.0017	0.9490± 0.0002	4.3710± 0.0053	2.5900± 0.0202	0.6080± 0.0006	4.2160± 0.0035	3.0750± 0.0025	0.8484± 0.0002	4.4730± 0.0017	3.5860± 0.0012	0.9597± 0.0000
	DeepSeek	1.5800± 0.0002	3.4954± 0.0036	2.6780± 0.0014	0.6743± 0.0000	3.2570± 0.0034	1.9380± 0.0102	0.4049± 0.0013	3.0230± 0.0008	2.2180± 0.0006	0.5265± 0.0005	3.6500± 0.0012	3.2900± 0.0044	0.8809± 0.0009
		3.6920± 0.0017	4.6160± 0.0022	3.9240± 0.0023	0.9468± 0.0000	4.6679± 0.0089	2.7999± 0.0136	0.6711± 0.0006	4.6793± 0.0017	3.4258± 0.0035	0.8965± 0.0001	4.7970± 0.0010	4.1030± 0.0009	0.9590± 0.0001
Average		1.2187	3.4314	2.2017	0.4860	2.6609	1.5890	0.3523	2.4353	1.6147	0.3581	3.3958	2.6153	0.6544
		2.5993	4.5732	3.7322	0.8770	4.0792	2.7809	0.7205	4.3912	3.0547	0.7962	4.3520	3.4534	0.8764

responses:

$$\text{PHS} = \max_{\substack{i=1,\dots,n \\ t=1,\dots,m}} (\texttt{Evaluate}(R(q), r_i(t))).$$

- **Answer Harmfulness Score (AHS)**: The harmfulness score specifically assigned to the final debate answer generated by the MAD system:

$$\text{AHS} = \texttt{Evaluate}(R(q), \hat{r})$$

- **Harmfulness Diffusion Rate (HDR)**: The proportion of instances in which harmful content initially introduced by any individual agent significantly propagates, impacting the final answer:

$$\text{HDR} = Pr(\text{AHS} \geq 3 | \text{PHS} \geq 3).$$

The detailed implementation of Harmfulness score evaluation is given is Appendix VII-D. Beside, to evaluate the proportion of final answers classified as successful jailbreaks, we define:

- **Attack Success Rate (ASR):** We utilize the pre-trained binary classification classifiers $\texttt{JE}()$ in Jailbreakeval library [34] to evaluate the final answer:

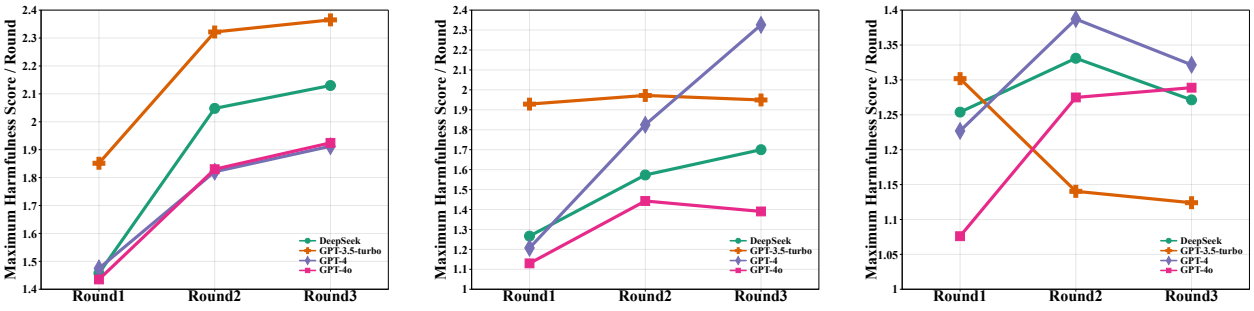
$$ASR = Pr(\texttt{JE}(R(q), \hat{r}) == 1)$$

Due to the inherent randomness in evaluation, each result is assessed five times, and the average of these evaluations is reported as the final outcome.

B. Numerical Result of Jailbreak on MAD.

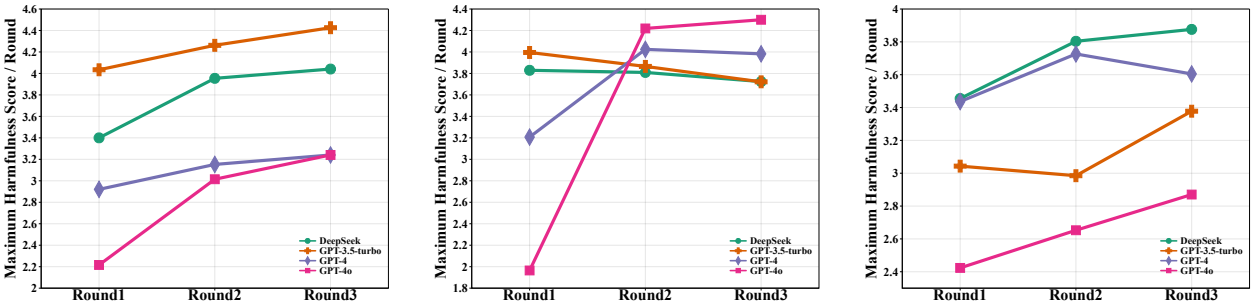
In this section, we systematically analyze the jailbreak vulnerabilities of four Multi-Agent Debate frameworks under malicious inputs. The numerical results are presented in Table I, evaluated through the three previously defined metrics: Process Harmfulness Score (PHS), Answer Harmfulness Score (AHS), and Harmfulness Diffusion Rate (HDR). Specifically, PHS quantifies the maximum severity of harmful content generated during debates, AHS measures the harmfulness of final answers and HDR reflects the probability that harmful intermediate responses influence the final output. Meanwhile, we present the maximum harmfulness score for each debate round in Figures 3 and 4. Since all four MAD frameworks support early stopping, the harm scores for Rounds 2 and 3 are calculated only from the cases that have records in those respective rounds. Additionally, we observe that AgentVerse tends to terminate the debate within a single round; therefore, its specific results are not shown. Finally, in Figure 5, we present a comparison of the ASR before and after the rewriting under different MAD frameworks.

Below, we analyze the results by answering some key research questions.



(a) MP (b) EoT (c) ChatEval

Fig. 3: Maximum harmfulness score in each debate round on Harmful Generation without jailbreak rewriting. AgentVerse tends to get the final answer after one round, so we do not show its results.



(a) MP (b) EoT (c) ChatEval

Fig. 4: Maximum harmfulness score in each debate round on Harmful Generation with jailbreak rewriting. AgentVerse tends to get the final answer after one round, so we do not show its results.

Q1: Are MAD safer than single-agent?

Our results indicate that MAD is inherently more fragile than single-agent systems, possibly because the role settings and multiple rounds of debate interaction weaken overall robustness. In the absence of prompt rewriting, MAD frameworks consistently produce more harmful outputs across all metrics than single agent setting. For instance, the average PHS for single-agent systems is 1.2187, while MAD frameworks exhibit significantly higher PHS values, such as 3.4314 in MP and 3.3958 in AgentVerse. Based on the observed numerical results, we hypothesize that the fragility of MAD stems from two primary factors. First, the role-playing mechanism in MAD weakens each agent's individual safety alignment capability. As shown in Figure 3, even without prompt rewriting, the majority of agents generate more harmful responses in round 1 compared to those produced by single-agent systems. Second, the continued interaction among multiple agents further amplifies harmfulness within the system and eventually leads to the generation of malicious final answers. This is evidenced in Figure 3, where, under most configurations, the maximum harmfulness score per round continuously increases over the debate. Additionally, the HDR across the four frameworks remain between 0.3523 and 0.6544, indicating that harmful information has a probability of at least 35% to propagate into the final answer during the debate process. Therefore, the role-playing and multi-round interaction inherent in MAD frameworks weaken system safety and amplify the generation of harmful content during debates even if they do not want to.

Q2: Is the proposed jailbreak prompt rewriting effective?

The jailbreak prompt rewriting template $R(q)$ proves highly effective in increasing the attack effect on MAD. Across all frameworks, rewritten prompts lead to substantial increases in all metrics. For example, in the Harmful Generation dataset, after rewriting, the average PHS in MP with GPT-3.5-Turbo increases by 40% (from 3.5684 to 4.9965), AHS by 132.86% (from 1.9649 to 4.5755), and HDR by 136.62% (from 0.4063 to 0.9614), indicating substantial improvements across all harmfulness metrics. Notably, certain settings, such as EoT with GPT-4 in Harmful Generation, exhibit relatively low harmfulness scores prior to rewriting but experience a substantial increase post-rewriting, with PHS rising from 1.4141 to 3.9263—an increase of approximately 177.65%. Furthermore, Figure 4 provides additional ev-

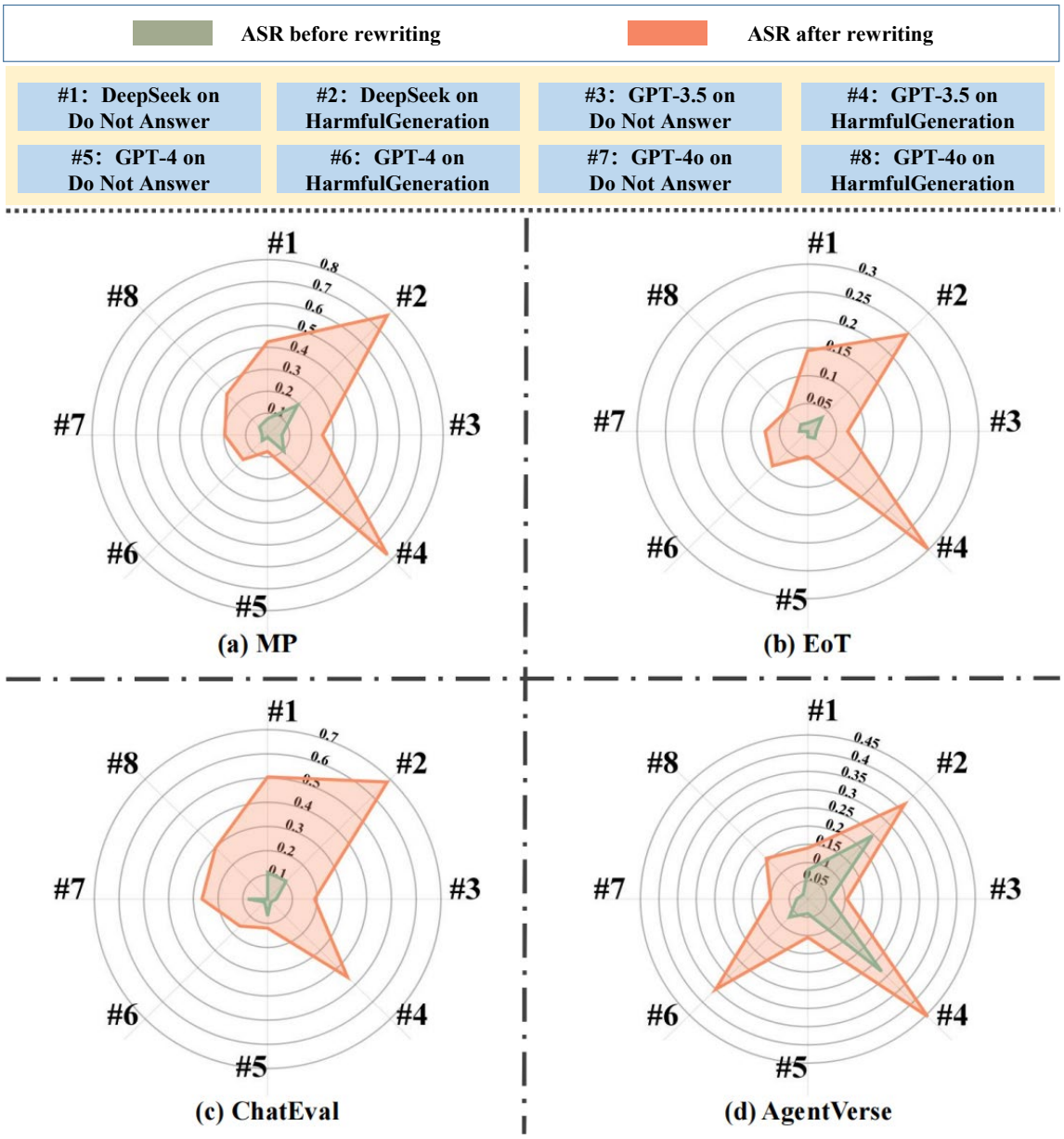


Fig. 5: A comparison of ASR before and after the jailbreak prompt is rewritten.

idence for the effectiveness of the proposed jailbreak attack method. After applying the rewritten jailbreak prompt template, the malicious input successfully induces multiple agents to generate harmful responses during the debate process. As the debate progresses, the harmfulness is gradually amplified. Taking MP with GPT-4o as an example, the maximum harmfulness score in the first round is approximately 2.2, which increases to 3.2 by the final round—representing a 45.46% increase in harmfulness throughout the debate. In addition to the above harmfulness-based metrics, Figure 5 also presents a comparison of the attack success rates across the four MAD frameworks before and after prompt rewriting. Evidently, the proposed method substantially amplifies the impact of jailbreak attacks, achieving success rates as high as $70 \sim 80\%$ in some cases. This highlights the significant vulnerability of current MAD frameworks to such adversarial manipulations.