Supplementary Materials for Content-free Logical Modification of Large Language Model by Disentangling and Modifying Logic Representation

Xin Wu[1,2], **Yuqi Bu**[1,2], **Yifei Chen**[1], **Yi Cai**[1,2*]

[1]South China University of Technology, Guangdong, China
[2]Key Laboratory of Big Data and Intelligent Robot (South China University of Technology) Ministry of Education
ycai@scut.edu.cn

	Validity		**Consistency**	
	Llama2	+LCF	Llama2	+LCF
Annotator1	0.52	1.38	1.92	1.90
Annotator2	0.88	1.68	1.96	1.96
Annotator3	0.76	1.60	1.84	1.76
Average	0.72	1.55	1.90	1.87

Table 1: Human evaluation.

Training Details

Given a natural language description as a premise, we can extract an invalid conclusion $Conclusion_{invalid}$ derived from this premise from the LFUD dataset. We then use GPT-3.5-turbo to generate a valid conclusion $Conclusion_{valid}$ derived from the same premise. From the LFUD dataset, we have constructed 540 such data pairs. Next, we identify identical tokens between $Conclusion_{invalid}$ and $Conclusion_{valid}$ (if none are found, they are skipped). For each pair of identical tokens, we extract their hidden representations from each layer of LLMs, including both attention and MLP layers. These extracted representations are used to train LCF for R_{input+} and R_{input-}. Since LLMs typically have 32 attention layers and 32 MLP layers, extracting all of them would result in excessive training data. In practice, we found that randomly sampling a proportion of the representations yields good results. Specifically, for each pair of identical tokens, we randomly select two layers between the 10th and 30th layers of attention and MLP to construct R_{input+} and R_{input-}. Finally, for Llama2, Llama3, Mistral, Vicuna, ChatGLM3, and Baichuan, we extracted 15,956, 13,400, 14,684, 15,956, 13,608, and 13,520 pairs of R_{input+} and R_{input-}, respectively. During training, the attentio and MLP representations from all layers are input into the same LCF for training. During inference, LCF only modifies the 10 attention or MLP layers with the highest distinctiveness. The distinctiveness of the l-th layer is calculated as follows: for all logically valid representations R^+ and logically invalid representations R^- in l-layer in the validation set, check whether all samples in R^+ are closer to the R^+ center than to the R^- center, and same as R^-.

*Corresponding author.

	Valid%(GPT-4)	**Δ Prob.**
2048,1024	96.56	6.29
1024,512	95.09	6.11
512,256	92.64	5.30

Table 2: Dimension Analysis

Human Evaluation

We manually evaluate the conclusions generated by LLMs before and after LCF from the following two dimensions: Logical Validity: Whether the conclusion can be logically derived from the premises. This includes three levels: invalid(0), partially valid(1), and fully valid(2).Content Consistency: Whether the content of the conclusion exceeds the scope covered by the premises. This includes three levels: completely unrelated(0), partially unrelated(1), and fully related(2). Three graduate students who are knowledgeable in logical fallacies and have good English proficiency scored 50 conclusions generated by Llama2 and 50 conclusions generated by Llama2+LCF. The specific comparison is shown in Table 1. From the results, it can be seen that LCF significantly improves the accuracy of Llama2. The average validity score is greater than 1.0, indicating that most of the generated conclusions are either partially valid or completely valid. In contrast, the original Llama2 has an average score of less than 1.0, suggesting that most conclusions are either invalid or partially valid. In terms of content consistency, Llama2 slightly outperforms Llama2+LCF. The reason is that Llama2+LCF usually generates longer content, and some annotators believe that it includes more words not present in the premise. However, in all results, no cases of generating completely irrelevant content were observed, indicating that the modification by LCF to improve logical validity does not affect content consistency.

Computation Efficiency

The parameter count of LCF is quite small, only requires 136MB parameters. Llama2 takes average 2.64 second to generate one conclusion, while Llama2 + LCF takes 2.94 second. Given the significant improvement LCF brings to logical validity, this additional time cost is entirely acceptable.

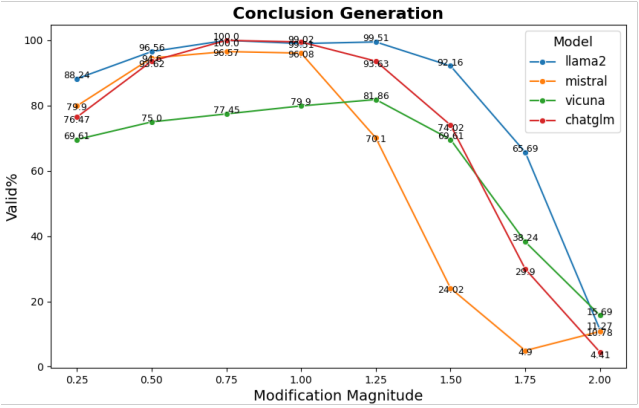

Figure 1: Modifying LLMs with different magnitude.

Dimension Analysis

We experiment with different dimension combinations in Table 2. The [2048, 1024] combination demonstrated the highest performance.

Modification Magnitude Analysis

We analyze the impact of magnitude on the validity of generated conclusions. Figure 1 demonstrates that magnitude significantly affects validity. When magnitude is below 1.0, increasing it improves validity. However, once magnitude exceeds 1.0, the validity of conclusions generated by LLMs declines sharply, dropping below 20% around a magnitude of 2.0. When magnitude exceeds 1.0, LLMs often produce incomplete and incoherent sentences, leading to this performance decline. This decline may result from substantial modifications disrupting the semantics of the representation.

Category Analysis

Figure 2 shows the validity distribution (in red) of conclusions generated by six LLMs across different types of fallacy data, as well as the validity distribution of conclusions after being modified by LCF (in blue). Without modification, LLMs can only generate valid conclusions for 60% of the data in most categories, especially in the ad hominem and deductive fallacy categories. After LCF modification, the proportion of valid conclusions generated across all categories significantly increases. However, the proportion remains relatively low for deductive fallacies, indicating that there is still room for improvement in the LLM's ability to perform deductive reasoning, which we will address in future work.

More Cases

Additional examples are shown in Tables 3, 4, 5, 6, 7, and 8. The conclusions generated by Llama2 are prone to invalid. After modification with LCF, the validity of the conclusions improves significantly. Interestingly, removing any loss used to train LCF does not significantly impact the validity of the generated conclusions. These partially trained LCFs can also enhance the validity of Llama2's conclusions. Conversely, when invalid modifications are applied, Llama2 generates invalid conclusions. These examples demonstrate that LCF editing of LLMs is both bidirectional and effective.

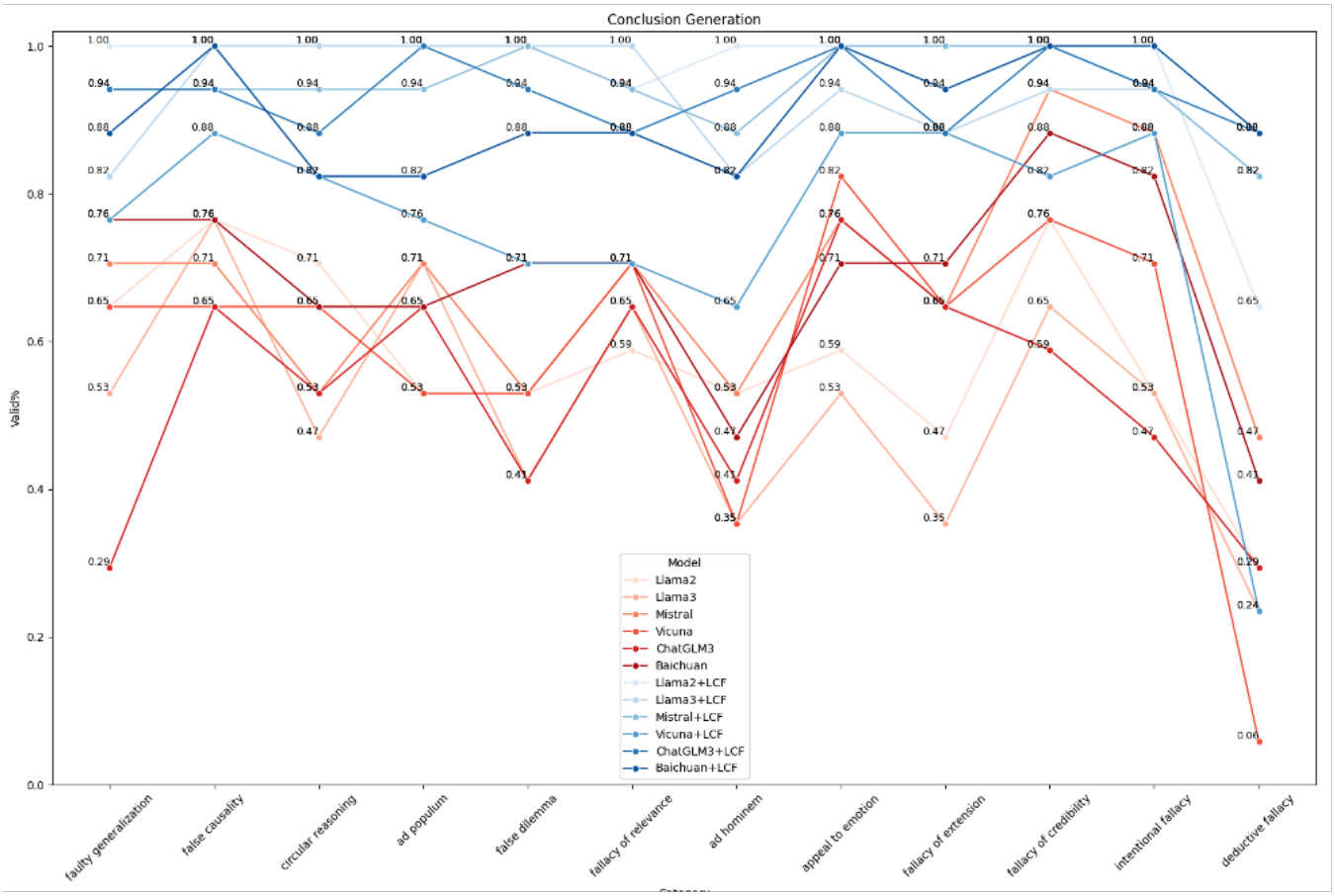

Figure 2: Category-wise Valid%

	Case 1	**Case 2**
Premise	John says that all flowers don't stay open forever. But John failed in botany.	Nobody can show me a flower that stays open forever.
Llama2	John's statement that all flowers don't stay open forever is incorrect.	Therefore, the idea of a flower that stays open forever is impossible.
Invalid Modification	John's statement that all flowers stay open forever is false.	No flower stays open forever.
Llama2+LCF	John's claim that all flowers don't stay open forever may not be accurate based on current scientific knowledge and understanding of plant biology.	Therefore, it is unlikely that there is a flower that stays open forever.
w/o L_{rec}	John's claim that all flowers don't stay open forever is likely not accurate based on his lack of expertise in botany.	It is unlikely that a flower will remain open indefinitely, as they are biological organisms that are subject to natural processes such as growth, development, and decay.
w/o $L_{content}$	John's statement about flowers is likely incorrect based on the principles of botany.	Therefore, it is unlikely that any plant can maintain an open state indefinitely.
w/o L_{logic}	John's statement that all flowers don't stay open forever may not be accurate based on his lack of expertise in botany.	Therefore, the statement "nobody can show me a flower that stays open forever" is likely true.

Table 3: After the modification by LCF, the validity of the conclusions is improved. In contrast, Llama2 or Llama2 with invalid modifications often generates invalid conclusions.

	Case 3	Case 4
Premise	Teacher A: Some students do not take their studies seriously.	If some of the residents of this city have a fever, then they are infected with COVID-19. A few people in the city are infected with COVID-19.
Llama2	Teacher B: Therefore, those students will not succeed in their academic careers.	Some of the residents of this city have a fever.
Invalid Modification	Students who don't take their studies seriously will fail.	Some of the residents of this city have a fever.
Llama2+LCF	Teacher B: This may be due to a lack of motivation or engagement in the learning process.	Therefore, some of the residents of the city may have a fever, but it is important to note that not everyone who has a fever is necessarily infected with COVID-19.
w/o L_{rec}	Teacher B: However, these students may excel in extracurricular activities, indicating a broader range of interests and abilities beyond academic performance.	Therefore, it is likely that some of the residents of this city have a fever and are infected with COVID-19.
w/o $L_{content}$	Based on this information, it is possible that some students in Teacher A's class may struggle with academic performance	Therefore, it is likely that some of the residents of this city have a fever and are infected with COVID-19.
w/o L_{logic}	Teacher B: But they are always the first to complain about their grades, which suggests they may be more invested in their education than they are letting on.	Some of the residents of this city may have a fever, but it does not necessarily mean that they are infected with COVID-19, as there could be other causes of fever in the city.

Table 4: After the modification by LCF, the validity of the conclusions is improved. In contrast, Llama2 or Llama2 with invalid modifications often generates invalid conclusions.

	Case 5	Case 6
Premise	Mary studied French in high school.	Mary studied French in high school and nobody has proven that she is not fluent in French.
Llama2	Mary is fluent in French.	Mary is fluent in French.
Invalid Modification	Mary is fluent in French.	Mary is fluent in French.
Llama2+LCF	Mary may have had more opportunities to practice speaking French if she had continued her studies beyond high school.	Mary's proficiency in French is likely due to her study of the language in high school, but it may not be the only factor.
w/o L_{rec}	Mary's ability to speak French fluently does not necessarily mean that she studied the language exclusively in high school.	Mary's proficiency in French is not necessarily a reflection of her high school studies, as there are many factors that can influence language proficiency.
w/o $L_{content}$	Mary may have learned basic vocabulary and grammar related to French.	Therefore, it is reasonable to assume that Mary is fluent in French.
w/o L_{logic}	Mary's ability to speak French does not necessarily mean she is fluent in the language.	Therefore, it is reasonable to assume that Mary is fluent in French, although her high school experience does not necessarily guarantee her proficiency in the language.

Table 5: After the modification by LCF, the validity of the conclusions is improved. In contrast, Llama2 or Llama2 with invalid modifications often generates invalid conclusions.

	Case 7	Case 8
Premise	Friend 1: Jenny has a collection of postcards.	People widely agree that Cynthia made a beautiful painting.
Llama2	Friend 2: Therefore, Jenny likes to travel.	The painting is a masterpiece.
Invalid Modification	Jenny loves traveling.	Cynthia is a talented artist.
Llama2+LCF	Friend 2: That's interesting, I have a collection of stamps.	The painting was likely well-received by those who saw it.
w/o L_{rec}	Friend 2: Therefore, Jenny likely enjoys traveling and collecting postcards as a hobby.	Cynthia's painting has been widely exhibited and received positive reviews.
w/o $L_{content}$	Based on this information, it is possible that Jenny enjoys traveling or has a interest in collecting souvenirs from different places.	Cynthia's painting is likely to be well-received by art enthusiasts.
w/o L_{logic}	Friend 2: Jenny's collection of postcards is likely not representative of her entire collection of possessions.	Cynthia's painting is likely to be appreciated by art enthusiasts, but it may not necessarily be a universal conclusion.

Table 6: After the modification by LCF, the validity of the conclusions is improved. In contrast, Llama2 or Llama2 with invalid modifications often generates invalid conclusions.

	Case 9	Case 10
Premise	X remains silent all the time.	Marie Kondo, a renowned organizing expert, agrees that I must organize the bedroom today.
Llama2	X is shy.	Marie Kondo's advice to organize the bedroom is sound.
Invalid Modification	X never speaks.	Marie Kondo thinks I should organize the bedroom.
Llama2+LCF	X may be shy or introverted, or they may have a valid reason for remaining silent in a particular situation.	Marie Kondo's endorsement of organizing the bedroom may motivate me to tackle this task.
w/o L_{rec}	X's silence may indicate a lack of confidence or shyness, but it does not necessarily mean that X is unable to speak.	Marie Kondo's endorsement of organizing the bedroom does not necessarily mean that it is a universal requirement or that it should be prioritized above other tasks.
w/o $L_{content}$	X may be shy or introverted, but it is also possible that X simply prefers to communicate in a more reserved manner.	Based on Marie Kondo's expertise, it is likely that organizing the bedroom will lead to a more tidy and peaceful sleeping space.
w/o L_{logic}	X's silence may indicate shyness or introversion, but it does not necessarily mean that X is not speaking, as there could be various reasons for their silence.	Marie Kondo's endorsement does not necessarily mean that organizing the bedroom today is the only thing that needs to be done.

Table 7: After the modification by LCF, the validity of the conclusions is improved. In contrast, Llama2 or Llama2 with invalid modifications often generates invalid conclusions.

	Case 11	**Case 12**
Premise	If my laptop was not at the office, then it would be stolen. My laptop was stolen.	Person A: The garden needs watering.
Llama2	Therefore, it must have been at the office.	Therefore, Person B should water the garden.
Invalid Modification	My laptop was stolen from the office.	Watering the garden is important.
Llama2+LCF	Therefore, it is likely that it was not at the office when it was stolen.	Person B: I can help with that, I have a watering can and some hoses that can be used to water the garden.
w/o L_{rec}	Therefore, it is likely that my laptop was not at the office, and it may have been stolen.	Person B: Therefore, it is likely that Person A has access to water and is able to water the garden, but it does not necessarily follow that Person A has the ability to water the garden.
w/o $L_{content}$	Therefore, it is likely that my laptop was not at the office.	Person B: Therefore, it is likely that the garden will remain healthy and thrive.
w/o L_{logic}	Therefore, my laptop was not at the office, but rather stolen, which contradicts the initial assumption.	Person B: The garden needs watering, but Person A is not available to water it, does not necessarily mean that the garden needs watering.

Table 8: After the modification by LCF, the validity of the conclusions is improved. In contrast, Llama2 or Llama2 with invalid modifications often generates invalid conclusions.