Google DeepMind

Fusion-Eval: Integrating Assistant Evaluators with LLMs

Lei Shu, Nevan Wichers, Liangchen Luo, Yun Zhu, Yinxiao Liu, Jindong Chen, Lei Meng

leishu@google.com

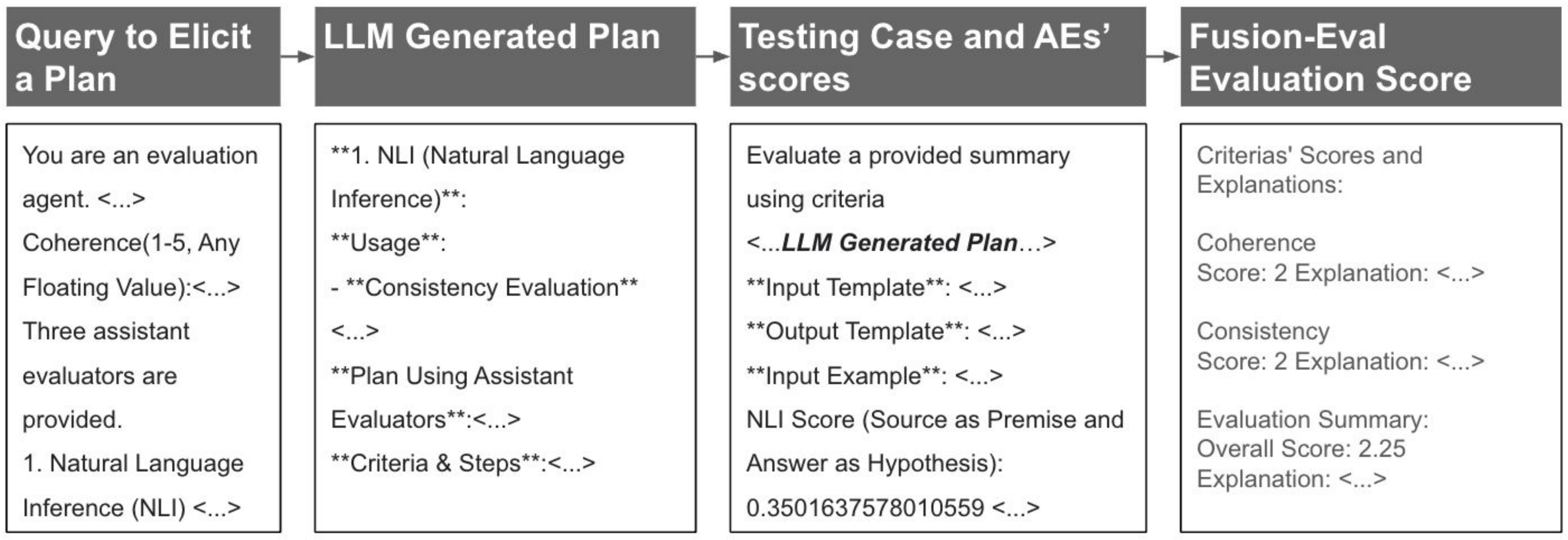
Research Problem

"Can Large Language Models (LLMs) integrate existing evaluators to achieve higher correlation with human judgments?"

Yes, Fusion-Eval!

Solution

Fusion-Eval is an innovative evaluation framework that integrates a variety of existing evaluators—termed ***assistant evaluators***—to enhance correlation with human judgment. Fusion-Eval prompts an LLM with an example to evaluate and scores given by assistant evaluators. In our work, we consider reference free evaluation. Fusion-Eval can evaluate any natural language task where assistant evaluators are available.



Query to Elicit a Plan

You are an evaluation agent. <...>
Coherence(1-5, Any Floating Value):<...>
Three assistant evaluators are provided.
1. Natural Language Inference (NLI) <...>

LLM Generated Plan

1. NLI (Natural Language Inference):
Usage:
- **Consistency Evaluation**
<...>
Plan Using Assistant Evaluators:<...>
Criteria & Steps:<...>

Testing Case and AEs' scores

Evaluate a provided summary using criteria
<...***LLM Generated Plan***...>
Input Template: <...>
Output Template: <...>
Input Example: <...>
NLI Score (Source as Premise and Answer as Hypothesis): 0.3501637578010559 <...>

Fusion-Eval Evaluation Score

Criterias' Scores and Explanations:

Coherence
Score: 2 Explanation: <...>

Consistency
Score: 2 Explanation: <...>

Evaluation Summary:
Overall Score: 2.25
Explanation: <...>

Figure 1: Workflow of Fusion-Eval with Plan (FE): Starting from the left, a query initiates the generation of a plan by the LLM. Once the plan is obtained, it is concatenated with the template. The template placeholders are filled in for each test example along with its specific assistant evaluators' scores. This complete prompt is then used to obtain the Fusion-Eval evaluation score from the LLM. A more detailed description of this workflow, including the prompt used, is provided in Appendix A.1.

Experiment

	Human Evaluation				
	Coh	Con	Flu	Rel	Overall
Reference-Based Metrics					
ROUGE-1	0.35	0.55	0.527	0.583	0.503
ROUGE-2	0.233	0.6	0.494	0.433	0.44
ROUGE-L	0.117	0.117	0.259	0.35	0.211
BLEU	0.217	0.05	0.326	0.383	0.244
CHRF	0.35	0.617	0.561	0.55	0.519
S1-CHRF	0.3	0.733	0.494	0.5	0.507
S2-CHRF	0.3	0.7	0.46	0.433	0.473
SL-CHRF	0.367	0.733	0.494	0.5	0.523
BERTScore	0.333	-0.03	0.142	0.2	0.161
MoverScore	0.217	-0.05	0.259	0.35	0.194
Source-dependent Metrics					
BARTScore	0.35	0.617	0.494	0.45	0.478
UniEval	0.683	0.75	0.661	0.667	0.728
DE-PaLM2	0.733	0.6	0.745	0.85	0.879
G-Eval (GPT-4)	0.733	0.583	0.778	0.883	0.912
Assistant Evaluators					
BLEURT	0.433	**0.767**	0.644	0.633	0.678
NLI	0.45	0.717	0.628	0.65	0.695
SumBLEURT	0.7	0.333	0.544	0.633	0.644
Aggregation of Assistant Evaluators (AE)					
AVG$_{(AE)}$	0.65	0.55	0.661	0.783	0.828
LLMSel$_{(AE)}$	0.7	0.75	-	0.767	-
CorrW$_{(AE)}$	0.667	0.65	0.678	0.783	0.845
Aggregation of AE and LLM Direct Evaluation					
AVG$_{(AE, DE-PaLM2)}$	0.717	0.583	0.728	0.85	0.895
AVG$_{(AE, G-Eval-GPT-4)}$	0.717	0.617	0.745	0.883	0.912
LLMSel$_{(AE, DE-PaLM2)}$	0.733	0.717	-	0.833	-
LLMSel$_{(AE, G-Eval-GPT-4)}$	0.733	0.717	-	0.85	-
CorrW$_{(AE, DE-PaLM2)}$	0.717	0.633	0.745	0.85	0.895
CorrW$_{(AE, G-Eval-GPT-4)}$	0.733	0.633	0.762	0.883	0.912
Fusion-Eval					
FE-PaLM2-NoPlan	0.767	0.617	0.728	0.867	0.895
FE-PaLM2	**0.783**	**0.767**	0.778	**0.917**	**0.962**
FE-GPT-4	**0.783**	0.762	**0.812**	0.9	0.946

Table 2: System-level Kendall-Tau (τ) correlations of different evaluators to human judgements on SummEval benchmark. The assistant evaluators, BLEURT, NLI and SumBLEURT, treat the article as a premise and the summary as a hypothesis.

	Human Evaluation					
	Coh (1-3)	Eng (1-3)	Nat (1-3)	Gro (0-1)	Und (0-1)	Overall (1-5)
Source-dependent Metrics						
UniEval	0.613	0.605	0.514	0.575	0.468	0.663
DE-PaLM2	0.669	0.688	0.542	0.602	0.493	0.66
G-Eval (GPT-4)	0.605	0.631	0.565	0.551	-	-
Assistant Evaluators						
BLEURT	0.316	0.461	0.384	0.638	0.432	0.464
PaLM2 Prob	0.583	0.606	0.637	0.441	0.676	0.687
Aggregation of Assistant Evaluators (AE)						
AVG$_{(AE)}$	0.556	0.637	0.626	0.579	0.672	0.697
LLMSel$_{(AE)}$	-	-	0.637	0.638	0.676	-
CorrW$_{(AE)}$	0.575	0.637	0.638	0.6	0.682	0.703
Aggregation of AE and LLM Direct Evaluation						
AVG$_{(AE, DE-PaLM2)}$	0.655	0.708	0.631	0.639	0.679	0.737
LLMSel$_{(AE, DE-PaLM2)}$	-	-	0.637	0.66	0.68	-
CorrW$_{(AE, DE-PaLM2)}$	0.666	0.711	0.641	0.65	**0.689**	0.742
Fusion-Eval						
FE-PaLM2-NoPlan	0.683	0.722	0.649	0.643	0.641	0.735
FE-PaLM2	**0.697**	0.728	0.651	**0.709**	0.632	0.764
FE-GPT-4	0.678	**0.747**	**0.691**	0.692	0.687	**0.774**

Table 3: Turn-level Spearman (ρ) correlations of different evaluators to human judgements on TopicalChat benchmark. BLEURT treats the fact and conversation as the premise and the response as the hypothesis. PaLM2 Prob represents the conditional probability of the response given the fact and conversation. The G-Eval scores for Und and Overall are missing because they aren't reported in their paper.

	SummEval					TopicalChat				
	Coh	Con	Flu	Rel		Coh	Eng	Nat	Gro	Und
BLEURT		✓		✓	BLEURT				✓	
NLI		✓			PaLM2 Prob			✓		✓
SumBLEURT	✓			✓						

Table 4: LLM-Suggested Assistant Evaluator Alignment for SummEval and TopicalChat Criteria. The criteria include coherence (Coh), consistency (Con), fluency (Flu), relevance (Rel), engagingness (Eng), naturalness (Nat), groundedness (Gro), and understandability (Und).

	FE-PaLM2				
	Coh	Con	Flu	Rel	Overall
BLEURT	0.583	0.867	0.733	0.65	0.717
NLI	0.6	0.783	0.75	0.667	0.733
SumBLEURT	0.75	0.467	0.633	0.717	0.683

Table 5: FE-PaLM2 and Assistant Evaluators System-level Kendall-Tau (τ) correlations on SummEval.

	FE-PaLM2					
	Coh	Eng	Nat	Gro	Und	Overall
BLEURT	0.524	0.558	0.59	0.662	0.622	0.67
PaLM2 Prob	0.711	0.784	0.808	0.588	0.711	0.792

Table 6: FE-PaLM2 and Assistant Evaluators Turn-level Spearman (ρ) correlations on TopicalChat.

	FE-GPT-4				
	Coh	Con	Flu	Rel	Overall
BLEURT	0.583	0.795	0.733	0.6	0.7
NLI	0.633	0.745	0.717	0.617	0.717
SumBLEURT	0.717	0.41	0.633	0.667	0.667

Table 7: FE-GPT-4 and Assistant Evaluators System-level Kendall-Tau (τ) correlations on SummEval.

	FE-GPT-4					
	Coh	Eng	Nat	Gro	Und	Overall
BLEURT	0.577	0.644	0.565	0.693	0.617	0.678
PaLM2 Prob	0.747	0.713	0.86	0.662	0.799	0.798

Table 8: FE-GPT-4 and Assistant Evaluators Turn-level Spearman (ρ) correlations on TopicalChat.

Conclusion

Fusion-Eval is an innovative aggregator using Large Language Models (LLMs) for diverse evaluation tasks. It effectively integrates assistant evaluators according to specific criteria. Empirical results show Fusion-Eval achieves higher correlations with human judgments than baselines. LLMs are very powerful, so it's interesting that augmenting LLMs with scores from simpler methods can improve performance in this case.