A Tale of Evaluating Factual Consistency: Case Study on Long Document Summarization Evaluation

Yang Zhong[1] and Diane Litman[1,2]

[1] Department of Computer Science, [2] Learning Research and Development Center
University of Pittsburgh, Pittsburgh, PA, 15213

Motivation

Challenge:

Ensuring factual consistency in long-document summarization is harder than for short texts due to:

- Nuanced inconsistencies spread across multiple sentences
- Binary labels masking subtle errors
- Rapid evolution in summarizers (from BART to GPT-4), yet outdated benchmarks still dominate.

Limitations of Prior Work

- Binary labeling often fails to capture partial inconsistencies in long summaries (100+ words)
- Prior work also lacks consistency in labeling schemes and reported evaluation metrics.

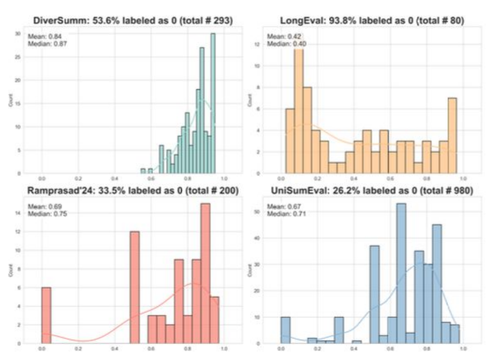

Figure 1: Distribution of summary-level continuous scores for summaries assigned the binary label of 0. For each dataset, we report the proportion of summaries containing errors (% labeled as 0) along with the total number of annotated article-summary pairs.

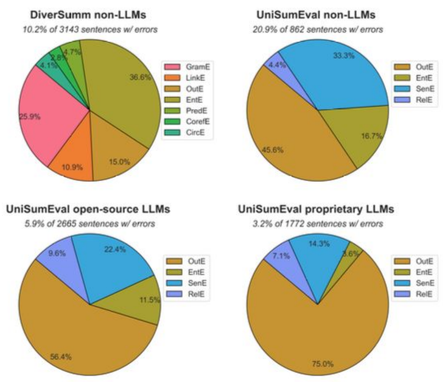

Figure 2: Error distribution across datasets and summarization models. Non-LLM summaries exhibit higher proportions of errors, of which entity (EntE) and out-of-article errors (OutE) are the most prevalent.

Research Questions

RQ1: Binary vs. continuous labels for summary-level evaluation?

RQ2: Do sentence-level predictions transfer to summary-level?

RQ3: How well do systems detect **fine-grained** errors?

RQ4: How do evaluation systems perform **as summarizers evolve** (from non-LLMs to LLMs)?

Experimental Setup

Datasets:

DIVERSUMM, LONGEVAL, RAMPRASAD'24, UNISUMEVAL

Labels: binary and continuous (at sentence and summary levels) Summaries generated by models from non-LLMs to GPT-4

Systems:

- Specialized fact-checkers: INFUSE, AlignScore, MC-FT5
- LLM-based: GPT4o, Gemini, BeSpoke, FineSurE
- Linguistic-informed: StructS-MC, StructS-BS

Result

RQ1: Binary and continuous evaluation settings exhibit different trends. Continuous summary-level scores offer a more comprehensive assessment than binary classification.

ID	Eval System	DiverSumm				Ramprasad'24				LongEval				UniSumEval			
	Ref. Label	Binary		Contin.		Binary		Contin.		Binary		Contin.		Binary		Contin.	
	Eval Metric	*AUC*	*bAcc*	r	ρ	*AUC*	*bAcc*	r	ρ	*AUC*	*bAcc*	r	ρ	*AUC*	*bAcc*	r	ρ
Specialized fact-checkers																	
1	INFUSE	87.15	59.31	0.55*	0.60*	74.16	60.86	0.36*	0.41*	**96.53**	**84.67**	0.72*	0.69*	52.91	50.86	0.05	0.04
2	AlignScore	84.59	**74.00**	0.53*	0.57*	75.55	54.46	0.42*	0.45*	91.20	76.00	0.74*	0.74*	57.41	53.74	0.21*	0.12*
3	MC-FT5	88.05	55.27	0.57*	0.62*	75.34	54.27	0.48*	0.45*	86.93	79.33	0.80*	0.77*	57.31	49.96	0.10*	0.12*
LLM-based																	
4	GPT4o	82.45	53.19	0.59*	0.59*	60.24	53.97	0.44*	0.33*	84.40	71.33	0.88*	0.87*	69.67	52.65	0.51*	0.41*
5	Gemini	83.11	55.10	0.52*	0.55*	56.58	54.35	**0.54***	0.33*	89.20	80.67	0.90*	0.86*	**71.18**	52.16	**0.54***	**0.47***
6	BeSpoke	61.92	53.21	0.35*	0.30*	66.91	53.58	0.49*	0.33*	90.13	68.67	**0.94***	**0.92***	69.61	53.29	0.41*	0.32*
7	FineSurE	81.36	68.37	0.49*	0.53*	62.63	51.86	0.32*	0.28*	84.93	**84.67**	0.86*	0.85*	69.24	**57.84**	0.41*	0.34*
Lingustic-informed																	
8	MC+rew.	88.62	61.27	0.59*	0.64*	**76.94**	54.23	0.49*	**0.47***	88.53	81.33	0.81*	0.76*	56.05	50.44	0.08	0.10
9	StructS-MC	**91.19**	62.54	**0.63***	**0.68***	71.68	58.08	0.43*	0.38*	89.60	80.67	0.90*	0.86*	54.91	51.30	0.08*	0.08*
10	BeSpoke+rew.	60.78	55.45	0.34*	0.29*	71.40	53.85	0.51*	0.40*	90.13	70.67	0.93*	0.90*	64.81	55.94	0.32*	0.25*
11	StructS-BS	89.65	69.72	0.62*	0.66*	72.03	**61.63**	0.47*	0.40*	86.40	74.67	0.86*	0.85*	58.99	55.67	0.17*	0.15*
	System-level Eval Metrics Ranking Correlation																
	AUC vs. bACC	$\rho = 0.45, \tau = 0.35$				$\rho = 0.41, \tau = 0.24$				$\rho = -0.02, \tau = 0.00$				$\rho = 0.50, \tau = 0.27$			
	AUC vs. Pearson	$\rho = 0.89^*, \tau = 0.84^*$				$\rho = -0.18, \tau = -0.11$				$\rho = -0.09, \tau = -0.04$				$\rho = 0.99^*, \tau = 0.94^*$			
	BACC vs. Pearson	$\rho = 0.25, \tau = 0.18$				$\rho = -0.19, \tau = -0.15$				$\rho = -0.57, \tau = -0.45$				$\rho = 0.52, \tau = 0.28$			

RQ2: On datasets that are sufficiently large and contain summaries with varying degrees of factual inconsistency, there exist **significant correlations** between sent-level and summary-level evaluations.

ID	Eval System	DiverSumm		Ramprasad'24			LongEval			UniSumEval		
	Eval. Task	Sent-level	Summ-level	Sent-level	Summ-level		Sent-level	Summ-level		Sent-level	Summ-level	
	Eval. Metric	*AUC*	*bAcc*	*AUC*	*AUC*	*bAcc*	ρ	*AUC*	*bAcc*	r	*AUC*	*bAcc*	r
	System-level Eval Metrics Ranking Correlation												
	sent-AUC vs. sent-bAcc	$\rho = 0.54, \tau = 0.38$		$\rho = 0.25, \tau = 0.18$			$\rho = 0.70^*, \tau = 0.56^*$			$\rho = 0.81^*, \tau = 0.60^*$		
	sent-AUC vs. summ-metric	$\rho = -0.11, \tau = -0.05$		$\rho = 0.51, \tau = 0.38$			$\rho = 0.86^*, \tau = 0.69^*$			$\rho = 0.72^*, \tau = 0.54^*$		
	sent-bAcc vs. summ-metric	$\rho = -0.36, \tau = -0.24$		$\rho = 0.06, \tau = 0.02$			$\rho = 0.80^*, \tau = 0.65^*$			$\rho = 0.81^*, \tau = 0.61^*$		

RQ3: LLM-based models (e.g. FineSurE) have higher recall but more false positives (Table 5)

Error Cat. (error/total : 392/5299)	OutE (212)	SenE (103)	EntE (50)	RelE (27)
AlignScore (904)	42.5%	29.1%	46.0%	40.7%
FineSurE (1129)	64.6%*	85.4%*	84.0%*	66.7%*
BeSpoke+rew. (1179)	59.9%*	77.7%*	64.0%*	51.9%

Table 5: Recall analysis in assessing factual consistency on UniSumEval. We report the sentence count for each erroneous type. We also include the number of predicted error sentences for each model, which indicates high false positives. (*) indicates a significant difference in recall compared to AlignScore.

ID	Eval System	UniSumEval					
	Gen.Model	non-LLMs		open LLMs		close-LLMs	
	Eval Metric	r	ρ	r	ρ	r	ρ
Specialized fact-checkers							
1	INFUSE	0.27*	0.29*	0.19*	0.19*	0.06	0.04
2	AlignScore	0.40*	0.32*	0.30*	0.27*	-0.04	-0.06
3	MC-FT5	0.24*	0.30*	0.28*	0.28*	-0.02	-0.02
LLM-based							
4	GPT4o	0.54*	0.53*	0.43*	0.37*	0.02	0.02
5	Gemini	**0.59***	**0.60***	0.41*	**0.42***	**0.14***	**0.10***
6	BeSpoke	0.46*	0.43*	**0.47***	0.40*	0.01	0.05
7	FineSurE	0.47*	0.44*	0.42*	0.34*	0.07	0.10
Lingustic-informed							
8	MC+rew.	0.24*	0.28*	0.27*	0.28*	-0.02	-0.01
9	StructS-MC	0.28*	0.31*	0.25*	0.26*	-0.00	-0.02
10	BeSpoke+rew.	0.44*	0.41*	0.41*	0.34*	-0.00	0.02
11	StructS-BS	0.41*	0.40*	0.30*	0.29*	0.01	-0.00

Table 6: Summary level results on UniSumEval separated by the types of summarizer model. We report Pearson correlation r and Spearman correlation ρ under **summary-level**, with * indicating statistical significance. The best model performance per column is **bold**. Green, orange and red indicate the performance ranking intervals of the system for each column, corresponding to top (rank 1-3), middle (rank 4-6), and bottom (rank 7-11) tiers, respectively.

RQ4: System performance declines as summarizers evolve, with most models showing sharp drops on closed-source LLMs.

Takeaways and Actionable Insights

- The **main types of factual consistency errors** have s**hifted as summarizer models continue to improve**: surface-level / grammatical → contextual and nuanced one
- Future work should move beyond overall summary-level correlations and focus on pinpointing specific errors.
- Future models require deeper reasoning and contextual understanding to detect errors.