

Fine-Grained Absolute Evaluation of Utterance Naturalness in Japanese AIWolf Dialogue: An Agreement Analysis between Human Annotators and LLM Judges

Chaeyoung Sung

Graduate School of Artificial Intelligence and Science,
Rikkyo University, Japan
seongchae35@gmail.com

Ryosuke Sueyoshi

Nihon Kogakuin College, Japan
ai168.sueyoshi@gmail.com

Kentaro Hoshi

Waseda University Global Education Center, Japan
drhoshiken@aoni.waseda.jp

Miyu Ueki

Meiji University, Japan
sq678223072849@gmail.com

Masato Taki

Graduate School of Artificial Intelligence and Science,
Rikkyo University, Japan
5064366@rikkyo.ac.jp

Abstract

In the AIWolf contest, agents are ranked relative to the other agents within a single game. Such within-game relative scores make it difficult to compare across years once the overall level of the participating agents changes. In this work we design a fine-grained evaluation framework that scores the surface naturalness of a single utterance on an absolute scale that does not depend on the year or on the other agents in the same game, using criteria defined over observable features of the utterance text. To examine how stable the framework is across the parties that apply it, we sampled 150 utterances from the 2025 and 2026 contest logs of the Japanese five-player track and had three human annotators and three LLM judges score them under the same criteria document, and we then compared human–human, human–LLM and LLM–LLM agreement. Agreement between humans and LLMs was not consistently high; however, agreement among the human annotators varied to a comparable degree, and for some criteria the agreement between an annotator and the LLM judges exceeded the agreement between any pair of human annotators. Agreement among the LLM judges was also higher than agreement among the human annotators for some criteria. These results indicate that, when LLM judges are used for absolute evaluation of AIWolf dialogue, their validity should not be assessed solely by agreement with one particular human annotator; the variation among human annotators has to be reported alongside it.

1 Introduction

In the natural language division of the AIWolf contest, AI agents play the werewolf game by conversing in natural language. When the quality of the generated utterances is assessed only through relative rankings within a game, a change in the overall level of the participating agents makes direct comparison with previous years difficult. An absolute evaluation framework is therefore needed, one that scores individual utterances against a common standard independently of the year and of the other agents in the same game.

To address this, we design a fine-grained evaluation framework that decomposes the surface naturalness of a single utterance into several observable features. LLM-as-a-Judge, in which an LLM scores generated text according to explicitly stated criteria and judging procedures, has recently been proposed (Liu et al., 2023; Ye et al., 2024; Kim et al., 2024). We apply this idea to absolute evaluation of AIWolf dialogue, in a form where both human annotators and LLM judges score utterances from one and the same criteria document.

For the framework to serve as an absolute scale, however, we first have to establish how consistently different parties judge when they are given the same criteria. We therefore had three human annotators and three LLM judges score an identical set of 150 utterances drawn from the 2025 and 2026 contest logs, and we compare human–

human, human–LLM and LLM–LLM agreement using the same set of metrics. Rather than treating the human scores as a single ground truth, we also measure the variation among the human annotators, so that a discrepancy between humans and LLMs can be examined for whether it is specific to the LLM judges or whether it reflects variation that arises generally when the criteria are applied.

The aim of this work is to build a fine-grained evaluation framework that allows utterance naturalness in AIWolf dialogue to be treated as an absolute scale comparable across years, and to examine how stable that framework is across the parties that apply it.

2 Related Work

For automatic evaluation with LLMs, several methods score generated text according to explicitly stated criteria and procedures: G-Eval (Liu et al., 2023), FLASK (Ye et al., 2024), which uses multiple fine-grained skill axes, and Prometheus (Kim et al., 2024), which supplies task-specific criteria. LLM-Rubric calibrates LLM predictions against a hand-designed multidimensional rubric while accounting for annotator-specific tendencies, and thus treats differences between human annotators explicitly (Hashemi et al., 2024). At the same time, LLM-as-a-Judge is known to exhibit biases such as position bias and self-preference (Zheng et al., 2023), so consistency among judges has to be examined as well as correspondence with human scores.

Human evaluation itself has been discussed in the NLG literature in terms of methodological variability and reproducibility. van der Lee et al. (2019) survey the substantial methodological variation in human evaluation for NLG and present literature-based best practices. Belz et al. (2023) attempted to verify the reproducibility of past human evaluations in NLP and report that fewer than 5% of the 116 papers remaining after filtering could be repeated from publicly available information alone. Amidei et al. (2019) survey 135 NLG papers and show that, judged against the established interpretation scales for inter-annotator agreement, most of the collected data fails to meet reliability thresholds; they argue that this reflects variation that is intrinsic to human interpretation of language, and recommend reporting agreement coefficients together with correlation coefficients. Following this, we do not decide agreement or disagreement from the absolute value of κ alone,

but report exact agreement rates, MAE, rank correlation and score distributions side by side, and emphasise relative comparison under identical conditions.

In the natural language division of the AIWolf contest, which is the target of this work, AI agents play the werewolf game by conversing in natural language (Kano et al., 2025). The generated dialogue contains questions, accusations, persuasion and rebuttals, so the naturalness of the utterances is an important object of evaluation when dialogue quality is considered. We focus not on game strategy or the correctness of the reasoning, but on how naturally each individual utterance reads as a conversational turn.

Building on this work, we do not evaluate the general capability of LLM-as-a-Judge as such. Instead we build an absolute evaluation framework for utterance naturalness in AIWolf dialogue and examine the variation between parties that arises when the framework is applied. Concretely, we compare human–human, human–LLM and LLM–LLM agreement in parallel over the same utterances and the same criteria. By using three human annotators, we avoid taking a single annotator as the reference point and instead compare the range that inter-annotator agreement can take with the agreement obtained against the LLMs. This perspective is related to LLM-Rubric, which treats annotator-specific tendencies as a calibration target, but we perform no calibration and observe the differences between parties under a shared criteria document.

3 Evaluation Design

3.1 Target Construct

What we evaluate is the naturalness of the utterance as it is expressed: whether the target utterance reads naturally as a conversational turn, judged at the level of surface expression only. The unit of evaluation is a single utterance. Appropriateness as a response to preceding utterances, the flow of the discussion, the validity of the reasoning, the truth of the content, deception about one’s role, and the quality of the game strategy are all outside the scope of the evaluation. Grammatical correctness is also distinguished from naturalness: ellipsis, nominal-ending sentences, and compressed expressions that occur naturally in spoken Japanese are not penalised on those grounds alone.

Type	Criterion	Range	Description
Primary	N1 Word choice and collocation	0–2	Are the words chosen, and combined with one another, naturally?
Primary	N2 Fit to conversational register	0–2	Is the register natural for a conversational or chat turn?
Primary	N3 Coherence as a single turn	0–2	Does the utterance hold together as one turn?
Auxiliary	X1 Mechanical repetition (within turn)	0/1/N/A	Is there unnecessary repetition of identical or similar expressions?
Auxiliary	X3 Orthographic noise	0/1	Are there mojibake, symbol runs, or spacing that impairs readability?
Auxiliary	F1 Use of colloquial forms	0/1	Are sentence-final particles, interjections, contractions, or trailing clauses present?
Reference	OVERALL Holistic impression	1–5	Naturalness of the whole utterance, judged without decomposition

Table 1: Evaluation criteria.

3.2 Evaluation Criteria

N1–N3 are the primary criteria and directly evaluate utterance naturalness. They form an ordinal scale from 0 to 2 and cover, respectively, word choice and collocation, fit to conversational register, and coherence as a single turn. X1, X3 and F1 do not contribute to the naturalness score; they are auxiliary criteria that record formal features, namely within-turn repetition, orthographic noise, and the presence of colloquial forms.

For N1–N3, a score of 2 means the utterance reads naturally with nothing jarring, 1 means the meaning is recoverable but part of the utterance is awkward, and 0 means there is a clear unnaturalness such that the utterance can hardly be called natural with respect to that criterion. X1, X3 and F1 are neither added to nor subtracted from the primary scores; each is recorded as an independent auxiliary criterion.

Separately from these, the human annotators also assigned OVERALL. OVERALL asks how natural the utterance felt as a whole, without decomposition into criteria, on a five-point scale from 1 (least natural) to 5 (most natural). It is not a substitute for N1–N3; it serves as a reference for checking how far the decomposed criteria correspond to the holistic impression, and how far that holistic impression is itself shared between annotators. OVERALL was not assigned by the LLM judges; only the human annotators provided it.

3.3 Criterion Selection

Candidate criteria were obtained by giving the same meta-prompt to three LLMs and having each produce two proposals, for a total of six. The meta-prompt restricted the object of evaluation to “the quality of expression of a single utterance in itself” and excluded appropriateness as a response

to preceding utterances, the flow of the discussion, and the validity of reasoning and strategy. It also required that grammatical correctness be distinguished from naturalness, and that the criteria be decomposable into concrete features observable in the utterance text rather than holistic impressions such as “human-like” or “feels natural”. The authors compared the six proposals in terms of fit to the target construct, clarity of the judging conditions, and likelihood of occurrence in the Japanese track, recorded the reasons for inclusion and exclusion, and fixed the evaluation specification. The prompt used at run time was then generated from the fixed specification by DeepSeek, which was forbidden to alter the criteria, the score ranges, the judging conditions, the treatment of N/A, or the reference scope. What the LLM generated at this stage was the surface form of the specification, not the content of the criteria. To keep the comparison conditions identical across models, this prompt was fixed and the same one was used for all three models; no per-model prompt optimisation was performed.

During selection, we excluded from the main scale the formulaic label notation and bulleted layout derived from the in-game protocol, sentence-final truncation, and repetition of syntactic templates. Label notation almost never occurred in the Japanese track, and sentence-final truncation contained more stylistic variation such as omitted full stops than genuine truncation. Repetition of syntactic templates (X2) requires reference to several utterances by the same speaker and therefore differs in scope from our construct, whose unit of evaluation is a single utterance. Criterion numbers are kept as they were at selection time, so X2 is absent from the final set.

Year	Retrieved	Used	Observations (game $\times$ player)	Non-null utterances
2025	285	180	900	6,872
2026	270	270	1,350	11,519

Table 2: Corpus statistics.

4 Experimental Setup

4.1 Corpus

For 2025 we selected the logs to use with the same success criterion as the script distributed by the organisers, together with the presence of a result record. For 2026 we obtained logs to which the success criterion had already been applied, so applying the same criterion excluded nothing and all 270/270 games were used. Nine teams participated in 2025 and 18 in 2026, of which three teams took part in both years. Utterance length differs between the years: the median character count fell from 86 to 40, and the median number of sentences from 4 to 2.

4.2 Evaluation Sample

The evaluation set consists of 150 utterances, 75 sampled from each of 2025 and 2026, all from the Japanese five-player track. At most one utterance was drawn per observation, and sampling was random with stratification by team. The three human annotators and the three LLM judges all evaluated the same 150 utterances. The human annotators were not shown the year, the team name, the character name or the LLM scores, and the presentation order was randomised separately for each annotator.

4.3 Evaluators and Analysis Metrics

There were three human annotators. Annotator 1 is an author of this paper, a non-native speaker of Japanese enrolled in a Japanese graduate school. Annotator 2 is a graduate student at another university and a co-author of this paper; they study AIWolf agents and are therefore familiar with the course of the game. Annotator 3 is a faculty member, is not an author of this paper, and participated only in the annotation. Annotators 2 and 3 are native speakers of Japanese. Annotator 2 is affiliated with a team that participated in the contest, so utterances from that team may be present in the evaluated corpus; however, the year, the team name and the character name were withheld during annotation, so no information that would identify their own team’s utterances was available

at the time of judging. None of the three annotators were informed of the analysis results or of the LLM judgments at the time of annotation. The LLM judges were DeepSeek, GPT-5.6-luna and Qwen3.6-27B.

The following metrics are used. The exact agreement rate is the proportion of utterances to which the two parties assigned the same value. The mean error (ME) is the mean of the score differences between the two parties; for human–LLM comparisons it is computed as (LLM judge score) $-$ (human annotator score), so a positive value means the LLM judge assigned the higher score and a negative value means it assigned the lower score. The mean absolute error (MAE) is the mean of the absolute values of those differences and expresses the magnitude of the discrepancy irrespective of direction.

N1–N3 form an ordinal scale from 0 to 2. As Amidei et al. (2019) point out, the absolute value of κ in rating tasks should not be interpreted through generic thresholds alone. Moreover, when the score distribution is skewed, the exact agreement rate and κ are affected differently. We therefore do not judge the absolute level of agreement from any single metric, but report the per-utterance exact agreement rate, the mean absolute error (MAE), linearly weighted Cohen’s κ (κ_{lin}) and Spearman’s rank correlation coefficient (ρ) side by side, applying the same metrics to human–human, human–LLM and LLM–LLM comparisons for relative comparison.

Note, however, that N1–N3 take three values while X1, X3 and F1 take two, so their κ values lie on different scales. We report values per criterion, but we do not compare κ across scales in order to conclude that one criterion is easier to agree on than another.

N/A in X1 indicates that there was no opportunity to judge repetition, so it is excluded from the agreement rate and from Cohen’s κ , and each comparison covers only those utterances to which both parties assigned 0 or 1. The mean value of X1 likewise excludes N/A and is computed over the utterances assigned 0 or 1.

Model	Model ID	Access	temperature	reasoning	max_tokens
DeepSeek	deepseek/deepseek-v4-flash-0731	OpenRouter	0	effort: low	32,000
GPT-5.6-luna	openai/gpt-5.6-luna	OpenRouter	0	effort: low	32,000
Qwen3.6-27B	qwen/qwen3.6-27b	OpenRouter	0	effort: low	32,000

Table 3: Execution settings for the LLM judges.

Criterion	1 & 2	1 & 3	2 & 3
N1	.553 / .513 / .016	.607 / .453 / .102	.700 / .327 / .155
N2	.880 / .167 / .050	.667 / .387 / -.020	.727 / .287 / .010
N3	.500 / .587 / .024	.500 / .587 / .049	.853 / .147 / .247
OVERALL	.333 / 1.100 / .086	.280 / 1.100 / .045	.407 / .773 / .148

Table 4: Agreement between human annotators. Each cell gives the exact agreement rate / MAE / κ_{lin} for the annotator pair.

4.4 Execution Settings for the LLM Judges

All three LLM judges were run through the OpenRouter API. At judging time the temperature was 0, the reasoning effort was low, and max_tokens was 32,000. top-p was not specified and the provider default was used. Up to three retries were issued when an output error occurred. When generating the candidate criteria, the temperature was 0.2 for all three models, and the temperature was also 0.2 when DeepSeek generated the run-time prompt.

5 Results

5.1 Distributions and Inter-Annotator Variation

Table 4 gives, for each pair of the three human annotators, the exact agreement rate, MAE and κ_{lin} , in that order. Table 5 gives the score distribution of each annotator over the primary criteria as counts of 0 / 1 / 2. The distributions differ by criterion and by annotator. For N2 in particular, Annotator 2 assigned 2 to 147 of the 150 utterances, whereas Annotator 3 assigned 2 to 111, 1 to 36 and 0 to 3. For N3 the distributions of Annotators 2 and 3 are close: their exact agreement rate is 0.853, MAE 0.147 and κ_{lin} 0.247. By contrast, for the N3 pairs involving Annotator 1 the exact agreement rate is 0.500 and MAE 0.587, so agreement among the human annotators itself varies substantially by pair. We therefore do not summarise inter-annotator agreement by a single value but treat it as a range over several pairs.

The rank correlations behave in the same way. Spearman’s ρ for the human pairs ranges from -0.003 to 0.182 for N1, from -0.015 to 0.092 for N2, from 0.076 to 0.290 for N3, and from 0.114 to 0.286 for OVERALL. If the annotators differed only in being systematically strict or lenient, the ranking would be preserved and the correlation would be high; in our data the ranking itself is not shared between annotators.

Criterion	Annotator 1	Annotator 2	Annotator 3
N1	9 / 49 / 92	2 / 40 / 108	4 / 19 / 127
N2	7 / 10 / 133	0 / 3 / 147	3 / 36 / 111
N3	16 / 58 / 76	0 / 14 / 136	3 / 12 / 135

Table 5: Score distribution of each annotator over the primary criteria.

For X1, which records a formal feature, Cohen’s κ with N/A excluded is 0.494 between Annotators 1 and 2 ($n = 132$), 0.325 between Annotators 1 and 3 ($n = 137$), and 0.485 between Annotators 2 and 3 ($n = 131$). Since X1 is binary and N1–N3 take three values, these κ values lie on different scales and we do not compare them directly to argue about ease of agreement. For F1, the proportion of utterances scored 1 was 94.0% for Annotator 1, 0.7% for Annotator 2 and 78.7% for Annotator 3, so Annotator 2 alone shows a markedly different distribution. This large gap indicates that F1 was very likely not interpreted consistently across annotators. Because we carried out no practice round and no reconciliation of disagreements, however, we cannot determine from the present data whether the gap stems from a particular annotator misunderstanding the criterion or from

Criterion	Human pairs	A1 vs. LLMs	A2 vs. LLMs	A3 vs. LLMs
N1	.016–.155	.023–.194	.012–.231	.307–.399
N2	–.020–.050	.000–.166	–.023–.000	.000–.089
N3	.024–.247	.004–.057	.145–.340	.280–.437

Table 6: Range of linearly weighted Cohen’s κ between human annotators and between human annotators and LLM judges.

insufficient specificity in the definition of F1’s colloquial forms itself. We therefore do not read the F1 result as positive evidence about the reliability of inter-annotator agreement, but treat it as an example of the importance of specifying and sharing judging criteria concretely. As F1 is an auxiliary criterion that does not enter the naturalness score (N1–N3), this does not affect the naturalness evaluation or our main conclusions.

5.2 Human–Human versus Human–LLM Agreement

Table 6 gives the range of linearly weighted Cohen’s κ (κ_{lin}) for N1–N3, over the three human pairs and over the three models for each human annotator (A1–A3 denote Annotators 1–3). Because agreement among the human annotators itself spans a range, it is difficult to call the human–LLM agreement high or low in isolation. The κ_{lin} values between the LLMs and Annotators 1 and 2 overlap the range of the human pairs. By contrast, the κ_{lin} values between Annotator 3 and the three LLMs are 0.307–0.399 for N1 and 0.280–0.437 for N3, exceeding the range of the human pairs in both cases (N1: 0.016–0.155; N3: 0.024–0.247). The direction of the score differences is not uniform either. For N2, the ME against Annotator 2 is –0.020 for DeepSeek and –0.007 for GPT-5.6-luna, and for N1 the ME of GPT-5.6-luna against Annotator 3 is –0.040, so there are combinations in which the LLM assigns the lower score. The MAE ranges also overlap: MAE among the human annotators is 0.147–0.587 over N1–N3, against 0.020–0.617 for human–LLM comparisons. The rank correlations show the same pattern: ρ for the human pairs is –0.003–0.182 for N1 and 0.076–0.290 for N3, whereas ρ between Annotator 3 and the three LLMs is 0.361–0.439 for N1 and 0.371–0.484 for N3, again exceeding the range of the human pairs. Finally, κ_{lin} among the LLM judges is 0.157–0.337 for N1 and 0.243–0.375 for N3. For N1 this lies entirely above the range of the human pairs (0.016–0.155), while for N3 it overlaps that range only slightly at the upper end (human pairs 0.024–0.247). Taken together, at least on the pre-

sent data it is difficult to interpret the discrepancy between humans and LLMs as a tendency specific to the LLM judges.

5.3 Concentration of Scores at the High End

Criterion	DeepSeek	GPT-5.6-luna	Qwen3.6-27B
N1	1.873	1.780	1.899
N2	1.960	1.973	2.000
N3	1.953	1.940	1.946
X1	0.027	0.071	0.027
X3	0.007	0.007	0.000
F1	0.860	0.840	0.785

Table 7: Mean score per criterion for the LLM judges.

Table 7 gives the mean score per criterion; the mean for X1 excludes N/A. For N1–N3 the LLM judges assigned a score of 2 to 78–100% of the utterances, showing a concentration at the high end. Qwen3.6-27B scored every utterance 2 on N2 (mean 2.000). This tendency is not confined to the LLMs, however: Annotator 2 also assigned 2 to 147 of the 150 utterances on N2. A high exact agreement rate therefore does not by itself show that the parties discriminate between individual utterances in the same way. Among the auxiliary criteria, Cohen’s κ between the LLM judges on F1, which records the presence of colloquial forms, was 0.608–0.694 for all three model pairs.

5.4 Handling of N/A in X1

X1 takes three values, 0, 1 and N/A, where N/A was used when the utterance consisted of a single sentence in which no identical or similar expression occurred twice, so that there was no opportunity to judge repetition. N/A is therefore not treated as a judging category alongside 0 and 1, and is excluded from the agreement rate and from Cohen’s κ for X1. Each comparison covers only those utterances to which both parties assigned 0 or 1, so the effective sample size for X1 differs by pair: $n = 132$ for Annotators 1 and 2, $n = 137$ for Annotators 1 and 3, and $n = 131$ for Annotators 2 and 3. Pairwise deletion was applied in the same way between LLM judges. The number of N/A

assignments was 10, 12 and 11 for the human annotators, against 0 for DeepSeek, 24 for GPT-5.6-luna and 2 for Qwen3.6-27B. Even for this formal decision of whether an opportunity to judge exists, the human annotators produced similar counts while the LLM judges differed considerably.

5.5 Verification of Quoted Evidence

Model	With quote	Not in source	Rate
GPT-5.6-luna	182	3	1.6%
DeepSeek	164	5	3.0%
Qwen3.6-27B	133	18	13.5%

Table 8: Verification of quoted evidence against the source text.

We mechanically checked whether the quotations output as evidence for a judgment appear in the source utterance (Table 8). The proportion not found in the source was 1.6% for GPT-5.6-luna, 3.0% for DeepSeek and 13.5% for Qwen3.6-27B. Most mismatches were due to orthographic conversion, but Qwen3.6-27B also produced cases in which several matching spans were listed and the output deviated from the required format. This is attributable to our specification not defining how to quote when more than one span matches. Even with an identical fixed prompt, therefore, ambiguity in the output specification can induce different behaviour in different models.

6 Cases of Divergent Judgments

Inspecting the utterances on which the parties disagreed, we find cases containing surface awkwardness: broken usage such as 「見ますです」, abrupt shifts of register from assertion to imperative, and unnatural word order. For these utterances, Annotator 1 sometimes assigned a low score while the LLM judges assigned a high one. Conversely, utterances to which Annotator 1 assigned a high score also contained particle ellipsis and compressed expressions, so compression and the use of jargon do not by themselves determine unnaturalness.

Representative examples are the following.

“発言の整合性と票の動きを中心に見ますです”

(“I’ll focus mainly on the consistency of statements and the movement of votes.” The polite auxiliary is doubled, yielding an ill-formed ending.)

“アスカ真寄りだな。灰は今、判断材料を落とせ”

(“Asuka looks more likely to be genuine. Grey players, drop your evidence now.” The register shifts abruptly from assertion to imperative.)

“占い師3人は本物と狼と嘘つき!タクミが悪い狼さんだよ!”

(“Of the three seers, one is genuine, one is a werewolf and one is a liar! Takumi is the bad wolf!”)

“@ミナコ 方針なき後出しは偽の典型。公約通り最寡黙のアスカを占い黒。Over”

(“@Minako, coming out late with no stated policy is typical of a fake. As promised, I divined the quietest player, Asuka, and the result is black.”)

In one 203-character utterance that put separate questions to three different players within a single turn, the N3 scores were 2 from Annotator 1, 2 from Annotator 2, 1 from Annotator 3, 2 from DeepSeek, 1 from GPT-5.6-luna and 0 from Qwen3.6-27B. The human annotators split between 2 and 1, and the LLM judges covered all three values. Here the boundary between “holds together naturally as one turn” and “a patchwork of independent fragments” was not sufficiently clear for utterances of this form. Differences between parties can therefore arise not only from properties of the LLM judges but also from vagueness in the boundaries of the judging conditions.

7 Discussion

7.1 Comparison without Relying on a Single Metric

Rather than concluding that the parties do or do not agree from the absolute value of κ , we reported the exact agreement rate, MAE and κ_{lin} together and examined the relative relations between parties. Among the three human annotators, even for the same criterion N3 the exact agreement rate ranges from 0.500 to 0.853, MAE from 0.147 to 0.587 and κ_{lin} from 0.024 to 0.247, so agreement varies considerably by pair. This shows that if human evaluation is treated as a single reference point, the interpretation of agreement with the LLMs changes depending on which annotator is chosen.

Nor can it be said that making the criteria fine-grained by itself raised human agreement uniformly. For OVERALL, which is judged without decomposition into criteria, the exact agreement rate for the human pairs is 0.280–0.407, MAE 0.773–1.100 and κ_{lin} 0.045–0.148, showing the same kind of variation by pair as N1–N3.

7.2 Stability of the Framework and Inter-Annotator Variation

To use the absolute evaluation framework designed here in a stable way, it is necessary to examine not only agreement with a particular human annotator but the variation between parties itself. In our data, agreement among the human annotators spans a wide range, human–LLM agreement overlaps that range, and for several criteria the comparison with Annotator 3 exceeds the maximum over the human pairs. For some criteria, agreement among the LLM judges was also higher than agreement among the human annotators. Disagreement with a particular human annotator therefore cannot be read directly as a failure of the LLM judges. Just as LLM-Rubric (Hashemi et al., 2024) treats annotator-specific tendencies as a calibration target, assessing the validity of an absolute evaluation framework requires reporting how much variation there is among the human annotators and where the distance to the LLMs falls within it, rather than treating “the human” as a single reference value.

7.3 Interpreting the Year Difference

Annotator 1 tended to give lower scores in 2026 than in 2025 (OVERALL -0.827), Annotator 2 showed almost no difference (-0.067), and Annotator 3 in fact scored 2026 higher ($+0.227$). The year differences for the three LLM judges were small for all of N1–N3, but OVERALL was not assigned by the LLM judges and cannot be compared for that criterion. The year difference thus fails to agree between annotators not only in magnitude but in sign. In addition, the composition of participating teams, utterance length, and the shift in Annotator 1’s scoring between sessions cannot be separated from the year difference. In particular, Annotator 1’s OVERALL shifted by $+0.710$ between sessions, which is of the same order as the observed year difference of -0.827 . We therefore do not treat the year difference as a main result, but as a secondary observation that we are unable to interpret.

8 Conclusion

We designed a fine-grained absolute evaluation framework that decomposes the surface naturalness of a single utterance into several observable criteria, so that utterance naturalness in the AI-Wolf contest can be compared without depending

on the year or on relative standing within a game. We then had three human annotators and three LLM judges score the same 150 utterances under the same criteria, and examined how stable the framework is across parties by comparing human–human, human–LLM and LLM–LLM agreement.

Agreement between humans and LLMs was not consistently high; at the same time, agreement among the human annotators also varied considerably by pair and by criterion. For some criteria, agreement with the LLMs exceeded agreement between the human pairs, and agreement among the LLM judges was in some cases higher than among the human annotators. It is therefore difficult to interpret the human–LLM disagreement observed here as a problem specific to the LLM judges.

LLM judges may thus be usable for absolute, cross-year evaluation of AIWolf dialogue, but validating the resulting scores by agreement with a single human annotator is not sufficient. When the framework is used for absolute evaluation, the variation among human annotators, the clarity of the judging conditions for each criterion, and the consistency between parties all have to be checked together.

Limitations

- The target is restricted to the Japanese five-player track; generalisation to other languages, other numbers of players and other tasks is untested.
- The judging criteria were shared with the human annotators as a document, but no practice round and no reconciliation over disagreement cases were carried out. The results should therefore be read as holding under the condition that only the criteria document was shared.
- Two of the three human annotators are authors of this paper and were aware in advance of the aims of the study and of the intent behind the evaluation design. The results therefore do not directly indicate the agreement that a third party with no involvement in the framework would achieve given only the criteria document. Agreement remains low even when Annotator 3, who is not an author, is included, so the observed variation is unlikely to stem solely from involvement in the research. Annotator 1 is also a non-native speaker of Japanese and thus differs in linguistic background from the other two. These effects

cannot be separated from one another on the basis of three annotators.

- The evaluation set comprises 150 utterances (75 per year), so the scale per year is limited.
- The year difference cannot be separated from the composition of participating teams, utterance length, or shifts in an annotator’s own criteria.
- The run-time prompt generated by DeepSeek was used for all three models. The input conditions are thus identical across models, but DeepSeek alone judges using a prompt it generated itself. The effect of the prompt’s provenance on the evaluation results has not been isolated here.
- The positive rate for F1 differed extremely across annotators (94.0% / 0.7% / 78.7%), indicating that its judging condition was probably not interpreted consistently. Because no practice round or reconciliation was carried out, we cannot separate whether this stems from a particular annotator’s misunderstanding or from the under-specification of F1 itself. We therefore do not interpret F1 as evidence of reliable agreement. As F1 does not contribute to the naturalness score, this does not affect the primary results.
- The quotation format for F1 did not adequately specify what to output when several spans match, which caused malformed output from Qwen3.6-27B.
- The unit of evaluation is the utterance, but utterances originating from the same team are not statistically independent.
- Repetition of syntactic templates was placed outside the scope of single-utterance evaluation and is not part of the present scale.
- The primary criteria N1–N3 take three values while the auxiliary criteria X1, X3 and F1 take two, so their κ values lie on different scales and ease of agreement cannot be compared across criteria. Whether reducing judging conditions to observable features contributes to agreement needs to be re-examined with a design in which the scales are aligned.
- OVERALL was assigned only by the human annotators and cannot be compared with the LLM judges. How the LLM judges correspond to human evaluation on holistic impression is untested.

Ethical Considerations

The utterances evaluated in this work were generated by AI agents in a public contest and contain no personally identifiable information; the character names appearing in them are fictional names assigned by the game system.

The three human annotators took part voluntarily and were not compensated. Two of them are authors of this paper. The third was informed of the purpose of the annotation and consented to the use of their scores here. The task consisted of reading short utterances from game logs and posed no foreseeable risk.

One annotator is affiliated with a team that participated in the contest, so utterances from that team may be present in the evaluated sample. To limit the influence of this, the year, the team name and the character name were withheld during annotation and the presentation order was randomised.

We do not report scores broken down by team and make no claim about the relative quality of individual participating systems. The framework examined here proved unstable across the parties that apply it, and the scores it produces should not be used as the sole basis for ranking contest entries.

The LLM judges were accessed through a commercial API. Their outputs may reflect biases in their training data, and we make no claim that their judgments represent a general standard of naturalness.

Supplementary Material Statement

The game logs used in this work are published by the organisers of the AIWolf contest (2025: <http://133.167.32.100/aiwolf/2025/JSAI/>, 2026: <http://133.167.32.100/aiwolf/2026/JSAI/>, both accessed in June 2026). For 2026 we were provided by the organisers with logs to which the success criterion had already been applied, and we used those.

The list of identifiers for the 150 evaluated utterances, and the code used for utterance extraction, criterion generation, LLM judging and aggregation, are available at <https://doi.org/10.5281/zenodo.22975039>. The human annotation scores are released to the extent that the annotators consent. The game logs themselves are published by the organisers and are therefore not redistributed here.

The LLM judges were accessed through a third-party API, so exact reproduction of their outputs cannot be guaranteed.

References

- Jacopo Amidei, Paul Piwek, and Alistair Willis. 2019. Agreement is overrated: A plea for correlation to assess human evaluation reliability. In *Proceedings of the 12th International Conference on Natural Language Generation, Association for Computational Linguistics*, pages 344–354, Tokyo, Japan.
- Anya Belz, Craig Thomson, Ehud Reiter, and Simon Mille. 2023. Non-Repeatable Experiments and Non-Reproducible Results: The Reproducibility Crisis in Human Evaluation in NLP. In *Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics*, pages 3676–3687, Toronto, Canada.
- Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. LLM-Rubric: A Multidimensional, Calibrated Approach to Automated Evaluation of Natural Language Texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, pages 13806–13834, Bangkok, Thailand.
- Yoshinobu Kano, Neo Watanabe, Yuya Harada, Yuto Sahashi, Claus Aranha, Daisuke Katagami, Kei Harada, Michimasa Inaba, Takeshi Ito, Hirotaka Osawa, Takashi Otsuki, and Fujio Toriumi. 2025. [AIWolfDial 2025: Summary of Natural Language Division of 7th International AIWolf Contest](#). In *Proceedings of the 3rd International AIWolfDial Workshop, Association for Computational Linguistics*, pages 1–13, Hanoi, Vietnam.
- Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing Fine-Grained Evaluation Capability in Language Models. In *The Twelfth International Conference on Learning Representations*.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics*, pages 2511–2522, Singapore.
- Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. FLASK: Fine-grained Language Model Evaluation based on Alignment Skill Sets. In *The Twelfth International Conference on Learning Representations*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems, Volume 36*, pages 46595–46623.
- Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation, Association for Computational Linguistics*, pages 355–368, Tokyo, Japan.