Como Nossos Pais: Princípios Clássicos para a Avaliação de Sistemas de IA

Livy Real[1,2], **Altigran Soares da Silva**[2]

[1] Instituto Kunumi, Belo Horizonte, MG, Brasil
[2]Universidade Federal do Amazonas, Manaus, AM, Brasil
livy@kunumi.com, alti@icomp.ufam.edu.br

Abstract. *This work revisits epistemological and methodological foundations from several scientific fields, such as psychology, corpus linguistics, sociology, and neuroscience, in order to establish basic principles for the evaluation of systems based on Natural Language Processing and Artificial Intelligence. We present in detail the state of the art of evaluation in the field, discussing the reasons that lead to the need for automating system evaluation (the LLM-as-a-Judge paradigm). We introduce Five Basic Methodological Principles that should underpin the evaluation of AI-based systems, grounded both in theories and experiments from disciplines that have studied human perception for decades, as well as in empirical evidence from NLP itself.*

Resumo. *Este trabalho retoma bases epistemológicas e metodológicas de diversas áreas da ciência, como psicologia, linguística de corpus, sociologia e neurociência, para fundamentar princípios básicos da avaliação de sistemas baseados em Processamento de Linguagem Natural e Inteligência Artificial. Apresentamos em detalhe o estado da arte da avaliação na área, discutindo os motivos que nos levam à necessidade de automatizar a avaliação de sistemas (paradigma LLM-as-a-Judge). Introduzimos Cinco Princípios Metodológicos básicos que devem fundamentar a avaliação de sistemas baseados em IA, fundamentados tanto na teoria e experimentos destas áreas que já pesquisam a percepção humana há décadas quanto em evidências empíricas do próprio PLN.*

1. Introdução

Recentes transformações na área de Processamento de Linguagem Natural (PLN) e Inteligência Artificial (IA) passam a evidenciar a necessidade de avaliarmos os sistemas gerativos de forma mais ampla e efetiva. Ainda que motivada pela explosão da IA Gerativa (Generative AI, GenAI), a avaliação de sistemas baseados em IA e em PLN não é um tema novo. A dificuldade de avaliar as saídas de modelos gerativos [Novikova et al. 2017, Asli et al. 2020, Blagec et al. 2022], a compreensão de que muitas métricas clássicas, como BLEU e ROUGE, não eram as mais adequadas [Callison-Burch et al. 2006, Reiter and Belz 2009], a percepção de que estávamos simplificando tarefas de PLN como forma de aproximá-las de uma tarefa real [Kalouli et al. 2017, Marasović 2018], são temas conhecidos da comunidade científica. Apesar de conhecer as dificuldades da avaliação de sistemas baseados em IA, a comunidade científica não alterou o paradigma de avaliação na última década. Hoje, tarefas tradicionalmente delimitadas passaram a operar em cenários de 'mundo aberto', o potencial de geração de dados é virtualmente infinito e surgem os chamados 'comportamentos emergentes', fatores que intensificam a dificuldade de decidir **o que** avaliar.

Nesse contexto, a avaliação de sistemas baseados em IA não pode ser tratada apenas como um problema técnico de métricas ou automação, mas como um problema metodológico e epistemológico de captura do julgamento humano. Exploramos a ausência de princípios metodológicos consolidados para operacionalizar avaliações confiáveis, reproduzíveis e alinhadas às perspectivas humanas. Como contribuição, este trabalho: (i) sistematiza desafios atuais (2026) da avaliação em PLN/IA; (ii) articula fundamentos teóricos oriundos de áreas como psicologia cognitiva, linguística de corpus, sociologia e neurociência para discutir avaliação de sistemas de IA; (iii) propõe cinco princípios metodológicos básicos, apoiados na metodologia existente de anotação de dados, para a construção de *pipelines* de avaliação humana e/ou automatizada (Princípios da *Independência, Diversidade, Confiabilidade, Simplicidade e Clareza*); (iv) fundamenta estes Princípios com exemplos práticos da literatura em PLN/IA; e (v) discute como esses princípios podem orientar o desenvolvimento de avaliações mais robustas e alinhadas ao comportamento humano em cenários reais.

Diferentemente de trabalhos focados exclusivamente em métricas, *benchmarks* ou arquiteturas de avaliação automática, nosso objetivo é propor uma perspectiva teórica e metodológica integrada para avaliação de sistemas baseados em IA, centrada na operacionalização confiável do julgamento humano (ou *human-like*). Esse trabalho deve ser lido antes como um artigo de posicionamento (*position paper*) do que como um artigo experimental. O posicionamento investigado aqui baseia-se na sólida metodologia de anotação de *corpus* [Bird and Liberman 2001, Voutilainen 2012] e em sua relação direta com as dificuldades atuais de avaliação. Os princípios metodológicos da área, por sua vez, são diretamente correlacionados com teorias da captura de julgamento humano de outras áreas da ciência, como sociologia e psicologia.

2. Avaliação Automática e Avaliação Humana

Nos últimos anos, assistimos ao *boom* do paradigma *LLM-as-a-Judge* [Lin et al. 2021, Thakur et al. 2024]. É fácil compreender esse movimento à medida que observamos a quantidade de conteúdo que temos gerado e que deveríamos avaliar. Pode-se dizer sem exageros que temos capacidade para gerar infinitos novos conteúdos, já que não existe nenhum limitador ou regulação em relação a isso. Como e quando avaliar todo esse conteúdo são questões que permanecem abertas. Considerando a escala da adoção de sistemas baseados em IA por todos os setores (público, privado, corporativo) e por sujeitos individuais, é necessário que tenhamos uma forma de avaliação em escala. A avaliação em escala é também obrigatoriamente automática[1]. Em última análise, não há humanos suficientes, suficientemente rápidos, bem treinados, bem remunerados, com a expertise ideal, para proporcionar segurança no uso de produtos baseados em IA.

Dizer que a avaliação precisa ser automatizada não é dizer que não precisamos de humanos na avaliação [Wu et al. 2022]. Sem dúvidas, precisamos de humanos na avaliação e uma avaliação só será considerada *ground truth* se for conduzida por humanos[2]. Note que uma avaliação humana, um *ground truth*, não necessariamente é a melhor ou é perfeita, livre de erros [Foody 2024]. A perfeição não vem dos humanos. O que vem

[1]Ou, ao menos, uma parte dela deve ser automatizável.

[2]Há ainda tarefas sem tanta subjetividade, como *coding* e problemas matemáticos, nas quais, eventualmente, outras formas de avaliação podem ser tão válidas quanto o julgamento humana para tarefas que envolvam linguagem natural.

do humano é a veracidade, é o compromisso (*commitment*) de que um ser humano, natural, com suas capacidades intelectuais em funcionamento, analisou um dado artificial. Um *ground truth* não é necessariamente perfeito, mas é necessariamente humano. Ele é a fonte da verdade, a representação do que humanos daquela determinada conjuntura histórica entendiam como a verdade.

Além de termos a responsabilidade humana na criação de um *ground truth*, os avaliadores humanos também estão presentes em distintos momentos desses ciclos. Por exemplo, a regulamentação brasileira em vigor, seguindo a Resolução do Conselho Nacional de Justiça (CNJ) número 615/2025 [BRASIL. Conselho Nacional de Justiça 2025], obriga que o conteúdo gerado por IA no Judiciário seja revisado e validado por um ser humano, garantindo que a responsabilidade final por um conteúdo seja sempre de um juiz ou servidor. Em atividades da indústria, é comum realizar avaliações humanas em alguns pontos do desenvolvimento de um produto e uma grande avaliação humana antes de cada novo lançamento [Artemova et al. 2025]. Recentemente, também discute-se a necessidade da perspectiva humana no que diz respeito ao 'alinhamento com humanos' [Zhao et al. 2026]: a necessidade de que as IA estejam alinhadas a princípios éticos que protejam a vida humana.

No que diz respeito ao PLN, às avaliações textuais às quais estamos habituados, o que é a introdução do alinhamento humano? É justamente entendermos o que o humano (bem-intencionado) faria ou desejaria em determinada situação e considerá-lo como *ground truth*. O que **não** é alinhamento com o humano? É um ser humano simplesmente auditar uma resposta de LLM e dizer que ela não está errada. Um humano não encontrar erros em uma resposta gerada não é a mesma coisa de garantir alinhamento humano. Um assistente automático de saúde, por exemplo, não deve ser avaliado por não gerar erros factuais, mas sim, por reproduzir o que um médico/profissional de saúde faria em uma dada situação, suas estratégias conversacionais, escolhas de registro linguístico, vocabulário técnico, entre outros. Ao passo que a avaliação automatizada está se tornando um simples e rápido *check-list* que observa se uma LLM errou ou não, estamos perdendo o alinhamento com humanos. Estamos nos alinhando às máquinas e assumindo que outro humano que leia isso aceitará esse conteúdo como **possível** ou **plausível** [Jacovi and Goldberg 2020]. O conteúdo não reflete necessariamente o que um humano especialista faria, ele apenas não possui erros factuais. Aí está a questão imediata e prática do alinhamento com humanos para o PLN. A avaliação é onde conseguimos garantir que as máquinas estão alinhadas ao que um humano faria ou deseja que ela faça, e não que humanos estão gradativamente se alinhando ao conteúdo gerado pelas máquinas.

Considerando o irrevogável papel dos humanos, seja criando o *ground truth* de uma tarefa, seja fundamentando os princípios de alinhamento humano, seja garantindo a qualidade de produto, precisamos também considerar **como** operacionalizar essas avaliações. Se a maior parte da avaliação dos conteúdos gerados será feita por máquinas, não podemos deixar essa operacionalização ser feita de maneira que não reflita o que os humanos (especialistas, se necessário) fariam.

2.1. O Paradigma *LLM-as-a-Judge*

Como operacionalizar essa forma de avaliação automática é um dos temas mais relevantes da área atualmente. O paradigma conhecido como *LLM-as-a-Judge* [Thakur et al. 2024]

tem recebido bastante atenção e a pesquisa tem se desdobrado nesse sentido. Temos inúmeros *frameworks* científicos e da indústria já disponíveis [Zheng et al. 2023, Confident AI 2024, Lee et al. 2025, Li et al. 2025b] para uso. Estudos discutem desde estratégias estruturais, como o uso de várias LLMs como juízes, conhecido como *LLM-as-a-Jury* [Qian et al. 2026], até soluções técnicas finas, como o uso de técnicas de zero-, one- ou few-shot prompting nesse contexto [Raina et al. 2024, Jang and Silavong 2025] e o uso de LLMs juízes em tarefas multimodais [Chen et al. 2024]. Também foram propostas investigações sobre métricas gerais a serem avaliadas, como *corretude* ou *completude*, referências claras aos conhecidos 'acurácia' e 'recall' [Wei et al. 2025].

À medida que se percebeu que a automação da avaliação destas tarefas não significa simplesmente replicar as métricas antigas usando modelos de linguagem [Gao et al. 2025], a própria estrutura da avaliação passou a ser investigada. Aqui, temos trabalhos que envolvem pensar na avaliação como um problema de ranqueamento [Li et al. 2025a], como uma constante escolha entre duas possibilidades (*pairwise comparison*) [Thakur et al. 2024] ou como um esquema de comparação iterativa em estilo de 'torneio', no qual múltiplas respostas são confrontadas sequencialmente até a determinação de um ranking final [Sandan et al. 2025].

Abriu-se também espaço para fazer a avaliação considerando os novos desafios que a tecnologia permite. *Groundedness*, hoje uma grande questão, no passado inexistente, torna-se mais uma variável de avaliação [Trautmann et al. 2024]. Vieses (algo que sempre existiu, mas estava menos em evidência, talvez pelos usos mais restritos de nossos modelos 'antigos', menos 'mundo aberto') passam a ser também investigados e surgem repentinamente, talvez repentinamente demais, inúmeros datasets para controle de vieses [Pagano et al. 2023]. Reforça-se nesse contexto a avaliação de temas como nocividade (*harmfulness*) e toxicidade [Hartvigsen et al. 2022].

Porém, para qualquer pessoa interessada em fazer uma avaliação real e efetiva do seu próprio sistema, com suas particularidades, essas questões gerais são pistas ou indícios, mas não avaliam exatamente o(s) ponto(s) que deve(m) ser avaliado(s). Todos esses trabalhos, especialmente os grandes datasets especializados, são muito úteis, sobretudo para a comparação de grandes modelos de linguagem, que hoje são lançados quase diariamente. Ter essas ferramentas de avaliação, *leaderboards* que tentam mensurar quase tudo, ao mesmo tempo, ajuda muito a ranquear esses modelos, a identificar como está o estado da arte e deveriam nos guiar nas escolhas de quais modelos utilizar e por quê. No entanto, ao avaliar um dado sistema, deve-se considerar o que aquele domínio e aquela tarefa especificamente precisam para serem considerados adequados em seu contexto de uso. O trabalho de [Marinho et al. 2026], por exemplo, é a tentativa de criar métricas específicas para o domínio jurídico. Um critério geral de corretude não é mais suficiente; precisamos avaliar a qualidade de cada um dos aspectos que importam nesse sistema e em seu uso real.

3. A Avaliação Ideal

Vimos que (i) o momento tecnológico impõe novas formas de avaliação caso desejemos ter controle do que produzimos com IA; (ii) não é possível escalar a avaliação sem automatizá-la, ao menos em parte; (iii) cada sistema precisará de uma avaliação diferente, não há métricas universais a serem usadas para qualquer tarefa.

Ainda que para as tarefas de mundo aberto nas quais estamos interessados atualmente não haja métricas únicas, **como** capturar o julgamento de um sujeito tem sido investigado por diferentes áreas nas últimas décadas.

O próposito deste trabalho é retomar algumas práticas tradicionais, estabelecidas por outras áreas de pesquisa, que não o PLN, que guiam como se deve coletar os dados para capturar a inteligência/compreensão humana, especialmente os dados de avaliação, mas não só. A pesquisa sobre a inteligência já existe, nós somos a parte que automatiza, artificializa essa inteligência.

Para isso, apresentamos boas práticas metodológicas de avaliação baseadas em disciplinas como linguística de corpus, psicolinguística, neurociência e historiografia. Nós, da IA, precisamos aprender a avaliar a inteligência dos nossos sistemas de inteligência artificial. Já há muito conhecimento no que diz respeito a avaliar a inteligência em si. Devemos nos apropriar desse conhecimento. Essas recomendações não são exaustivas, mas trazem princípios básicos para assegurar uma avaliação justa, seja feita por humanos, seja por LLMs. A fundamentação desses princípios foi realizada a partir de leitura crítica e orientada da literatura em anotação de dados e artigos seminais de áreas correlatas, não representando revisão sistemática de nenhuma das áreas citadas.

4. Metodologia Básica para Avaliação de Sistemas Baseados em IA

Tentaremos abordar da forma mais didática possível princípios básicos consolidados em diversas áreas do conhecimento no que diz respeito à avaliação e metodologias para a captura isenta do julgamento humano. Exemplificamos a adoção de cada um desses princípios com base na literatura recente da área de PLN. Discutimos na seção 6, como estes princípios podem ser empiricamente validados.

4.1. Princípio da Independência

O Princípio da Independência diz respeito a cada juiz conduzir seu processo avaliativo independentemente dos demais; em termos práticos, trata-se de **avaliação às cegas**.

Nenhum dos avaliadores (sejam humanos ou LLMs) deve ver a resposta de outro avaliador [Wojcieszak and Price 2009] ou ter acesso à fonte do material avaliado (por exemplo, qual modelo gerou determinada resposta). Isso evita o Viés de Falso Consenso (*False Consensus Bias*) [Ross et al. 1977], que indica a tendência de juízes concordarem com outros ou com a resposta que acreditam que a maioria irá preferir. O trabalho de [Choi et al. 2025] investiga o Viés de Falso Consenso também entre LLMs-juízes.

O Princípio da Independência significa, por exemplo, que avaliadores humanos não devem ver as respostas de LLMs-juízes e concordar ou não com respostas já estabelecidas. Isso evita o *Acquiescence Bias*, conhecido na psicologia desde a década de 1960 [Messick and Jackson 1961], que caracteriza a tendência de seres humanos responderem positivamente a uma pesquisa ou avaliação. Evita também o Efeito de Ancoragem (*Anchoring Effect*), conhecido na psicologia cognitiva desde a década de 1970 [Tversky and Kahneman 1974], que caracteriza o comportamento humano por preferir julgamentos que retomam ('ancoram') uma informação inicialmente recebida. A percepção do Efeito de Ancoragem teve implicações práticas em áreas como economia, direito, design, ciência política, entre outras.

Considerando LLMs, o Princípio da Independência evita o Viés de Acordo (*Agreement Bias*), a tendência de uma LLM concordar com uma resposta previamente estipulada [Jain et al. 2025, Thakur et al. 2024]. Ajuda também a mitigar o Viés de Simpatia (*Sycophancy Bias*), a tendência de LLMs priorizarem agradar o usuário em detrimento de raciocínio independente [Fanous et al. 2025].

Tem sido prática recorrente humanos avaliadores partirem de respostas de LLMs-juízes, concordando ou não com etiquetas ou análises previamente apresentadas. O custo mental e psicológico de discordar de algo, especialmente se o avaliador humano deve explicar por que discorda, é muito maior do que o de simplesmente concordar [Gilbert 1991]. O juiz humano tende a discordar apenas quando a resposta da LLM for ilógica. Dessa forma, muitas avaliações de sistemas de PLN recentes não são confiáveis por priorizarem economizar tempo em vez de tomar o julgamento humano *a priori* como o verdadeiro *ground truth*. Esse tipo de metodologia exemplifica o alinhamento à máquina discutido anteriormente e não é confiável. Outros conceitos, como a Pressão dos Pares (*Peer Pressure*) [Durkin 1996], da sociologia, e a Conformidade (*Conformity*) [Beloff 1958], da psicologia social, são outras formas de vieses que devem ser mitigadas pela metodologia de avaliação. Curiosamente, já temos essa prática estabelecida nos processos de revisão por pares aos quais submetemos nossos trabalhos científicos.

O trabalho de [Choi et al. 2025] reproduz o experimento seminal na psicologia de [Ross et al. 1977] procurando entender se LLMs exibem o Viés de Falso Consenso, isto é, tendem a escolher a resposta que acreditam que a maioria escolheria. Os autores testam quatro modelos (OpenAI GPT-4, Anthropic Claude 3 Opus, Meta LLaMA-2 70B, Mistral AI Mixtral 8x7B) avaliando a resposta das LLMs e a resposta que as LLMs acreditam que será escolhida pela maioria. Todos os modelos possuem o viés. Para tentar mitigá-lo, os pesquisadores testam técnicas de *prompting*, oferecendo diferentes tipos de informação às LLMs relacionadas à tarefa. Os experimentos demonstram que oferecer no prompt informações em oposição aos julgamentos iniciais reduz o Viés de Falso Consenso.

[Jain et al. 2025] mede o Viés de Simpatia de LLMs-juízes ao avaliar códigos Python de alunos. O experimento, bastante robusto, conta com 66 submissões de código Python, 69 tarefas diferentes e 14 LLMs-juízes. Os modelos avaliadores devem avaliar o código recebido, com e sem uma avaliação prévia, que pode estar correta ou incorreta. Em números gerais, as LLMs-juízes conseguem capturar 60-80% dos erros de código quando não recebem avaliação prévia, porém quando recebem a informação prévia errada, esta taxa cai para 25%. Os experimentos de [Jain et al. 2025] conseguem quantificar quanto o julgamento das LLMs-juízes é afetado quando estas têm acesso a uma resposta, mesmo que errada, evidenciando fortemente a necessidade do Princípio da Independência.

4.2. Princípio da Diversidade

O Princípio da Diversidade diz respeito à tentativa de capturar compreensões e julgamentos de avaliadores com diferentes horizontes socioculturais e *backgrounds*. Na prática, trata-se de ter **múltiplos avaliadores** para uma tarefa.

Amplamente difundida na linguística de corpus, outra prática essencial para avaliação de sistemas de PLN/IA é ter pelo menos dois juízes diferentes na avaliação de um mesmo dado [Meyer 2002, McEnery and Hardie 2012]. Se há discordância entre

eles, um terceiro juiz deve ser acessado e contabiliza-se a maioria[3]. Em tarefas muito complexas, com muitas nuances, até 5 avaliadores (ou mais) podem participar de uma mesma tarefa. Por exemplo, em tarefas que envolvem semântica e pragmática, projetos com recursos para anotação, como o corpus SICK [Marelli et al. 2014] e o conhecido SNLI [Bowman et al. 2015], possuem no mínimo 5 anotadores humanos por instância e reportam não apenas o voto majoritário, mas todos os julgamentos realizados [Frenda et al. 2024].

Embora essa boa prática seja estabelecida na linguística de corpus e linguística descritiva considerando sujeitos humanos, para LLMs a prática também se sustenta [Verga et al. 2024], sendo um avanço recente da área de avaliação usar um júri e não apenas um juiz. [Verga et al. 2024] investiga o impacto de usar múltiplos LLMs-juízes em duas tarefas, avaliação de perguntas e respostas e avaliação de chatbots. Partindo de datasets conhecidos, como o TriviaQA e o ChatBot Arena, os dados foram reanotados por humanos altamente qualificados e o acordo das LLMs-juízes foi medido em relação a esse *ground truth*. Os autores investigaram LLMs de três famílias distintas para uso individual e conjunto. Nos cenários testados, o uso de LLMs em conjunto aumenta o alinhamento entre a avaliação automática e a humana, chegando a aumentar em 10 pontos a correlação de Pearson para o benchmark ChatBotArena Hard. Os autores sugerem que essa melhoria se dá pela mitigação tanto de vieses dos modelos, quanto de inconsistências nas execuções de um mesmo modelo.

Considerando os vieses intrínsecos dos modelos, [Zhu et al. 2026] propõem uma abordagem cíclica, com um rodízio entre as LLMs-juízes. Usando o MT-Bench e 5 LLMs-juízes (Qwen 2.5 7B Instruct, Llama 3.3 70B Instruct, GPT-5.2, Gemini 3 Flash e Claude Sonnet 4.6), os autores realizam uma decomposição da variância das pontuações atribuídas, mostrando que o efeito do juiz é a principal fonte de variabilidade, respondendo por mais de 90% da variância total observada no regime de avaliação com um único juiz escolhido aleatoriamente. A variância agregada cai para 27-40% com a abordagem de rodízio de juízes (cíclica) e, ao utilizar um júri com todos os LLMs-juízes testados, ela desaparece. Obviamente, o custo da avaliação pode se tornar proibitivo no cenário com todos os juízes, logo, os autores defendem o rodízio cíclico como um compromisso eficiente entre custo e confiabilidade na avaliação baseada em LLMs.

Ter múltiplos avaliadores de uma mesma instância, então, ajuda a mitigar vieses: sejam eles culturais, individuais (de modelo ou sujeitos humanos), além de reduzir erros de anotação e alucinações. No caso de avaliadores humanos, erros são esperados em qualquer tarefa, porém são mais pronunciados quando os avaliadores estão sob pressão, seja de tempo, seja financeira, seja por condições de trabalho insalubres [Snow et al. 2008, Fort et al. 2011, Martin et al. 2014].

4.3. Princípio da Confiabilidade

O Princípio da Confiabilidade considera a necessidade de mensurar quão efetivo e confiável é o próprio processo de avaliação. Na prática, trata-se de medir o **Acordo Inter-Anotadores enquanto** a tarefa de avaliação está sendo realizada.

Para garantir qualidade e consistência na avaliação, deve-se medir regularmente

[3]Há também uma vasta literatura, já em PLN, sobre a melhor forma de computar esse desacordo e sobre como os sistemas de IA deveriam tratá-los [Pavlick and Kwiatkowski 2019, Plank 2022]

o Acordo Inter-Anotadores (*inter-annotator agreement*, IAA) entre juízes humanos e/ou LLMs [Freitas 2024]. Uma pequena parte dos dados é avaliada por todos os avaliadores na mesma tarefa; isso garante confiabilidade e reprodutibilidade no processo, assegurando que todos os envolvidos estão aptos e compreenderam a tarefa. Métricas como Cohen's Kappa, Krippendorff's Alpha, Kendall's Tau ou Scott's Pi devem ser utilizadas conforme o tipo de tarefa [Artstein and Poesio 2008]. A porcentagem simples de acordo deve ser evitada, pois desconsidera o acordo ao acaso e o viés do anotador. Valores baixos de acordo indicam a necessidade de revisar critérios, instruções e o próprio processo de julgamento. A avaliação só deve ser considerada válida quando apresentar níveis de confiabilidade compatíveis com o padrão mínimo estabelecido.

[Carletta 1996] argumenta que o acordo não deve ser tratado como mero detalhe estatístico, mas como evidência de confiabilidade experimental. Estudos posteriores mostram que baixos níveis de acordo refletem problemas estruturais da própria tarefa, incluindo ambiguidades nas diretrizes, subjetividade excessiva, categorias mal definidas ou treinamento insuficiente de anotadores [Bayerl and Paul 2011]. Nesse contexto, um acordo baixo não deve ser interpretado como falha humana, mas como sinal metodológico relevante sobre a qualidade da avaliação.

Mais recentemente, [Abercrombie et al. 2025] argumentam que medidas de acordo também funcionam como mecanismos de controle de estabilidade temporal, mostrando que anotadores humanos podem apresentar inconsistências significativas ao longo do tempo mesmo em tarefas aparentemente simples. Os autores observam variações próximas de 25% em diferentes tarefas de PLN, sugerindo que a confiabilidade avaliativa precisa ser continuamente monitorada e não apenas assumida ou *a posteriori*.

Trabalhos consolidados em LLM-as-a-Judge que comparam o alinhamento de modelos juízes entre si ou com avaliadores humanos reportam suas métricas de acordo [Zheng et al. 2023, Thakur et al. 2024, Verga et al. 2024, Lee et al. 2025, Qian et al. 2026].

4.4. Princípio da Simplicidade

O Princípio da Simplicidade ressalta a necessidade de oferecermos as tarefas mais simples possíveis para o avaliador. Quanto mais simples a tarefa, mais fácil será termos um entendimento generalizado dela e um maior acordo entre os juízes [Yang et al. 2016]. Em termos práticos, na Era das LLMs, deve-se **evitar pedir as justificativas (*rationales*)** dos juízes [Madsen 2024], especialmente quando essa justificativa não tem um objetivo claro dentro do ciclo de desenvolvimento.

A Teoria da Carga Cognitiva (*Cognitive Load Theory*) [Sweller 1988], usada em educação, design instrucional e usabilidade, incentiva que tarefas evitem informação irrelevante ou complexidade desnecessária, dada a capacidade limitada da memória de trabalho humana [Chen et al. 2018]. Se uma tarefa impõe uma carga cognitiva muito alta, o desempenho e a aprendizagem são prejudicados [de Jong 2010].

Tendemos a supor que pedir uma explicação lógica para uma resposta garante que a avaliação seja correta, mas não é bem assim. Tanto pessoas quanto modelos de linguagem costumam criar justificativas que parecem convincentes para a primeira resposta que possuem. Os raciocínios expostos geralmente não refletem o processo real que levou à resposta. Isso confunde plausibilidade (*plausibility*) com fidedignidade (*faithful-*

ness) [Jacovi and Goldberg 2020]. Além disso, avaliadores humanos tendem a ser enviesados por explicações geradas automaticamente e aceitá-las mesmo que haja outra justificativa mais apropriada para a tarefa. Pedir justificativas detalhadas também é caro e demorado e não garante mais precisão [Madsen 2024]. Por isso, esse tipo de explicação só deve ser solicitado quando houver clareza sobre seu uso e quando fizer parte do ciclo de desenvolvimento do sistema.

Talvez esse seja o Princípio mais controverso nos dias de hoje, com tantos trabalhos na área perseguindo *chain-of-thoughts, rationales* e *explanations*. Porém, a Psicologia Cognitiva é clara quanto aos processos mentais implícitos no julgamento humano e trabalhos recentes investigando profundamente o uso de explicações no contexto de LLMs também já demonstram a ineficiência desse tipo de argumento *post hoc*.

[Fayyaz et al. 2024] investigam alinhamento humano e fidelidade de *rationales* produzidos por LLMs utilizando diferentes tarefas. Os autores comparam explicações geradas com métodos de atribuição baseados em gradientes e atenção interna do modelo buscando relacionar os raciocínios textuais ao que acontece internamente nos modelos. Os experimentos utilizam datasets com *rationales* humanos anotados e avaliam LLMs robustas (Llama-2, Llama-3, Mistral, GPT-3.5 e GPT-4-Turbo). A avaliação de alinhamento humano é realizada por meio de F1 entre *rationales* humanos e *rationales* gerados, enquanto a fidelidade é medida via *flip rate* (frequência com que a remoção das palavras consideradas importantes altera a decisão do modelo). Os resultados mostram que *rationales* produzidos apresentam maior plausibilidade discursiva, mas menor fidelidade ao processo decisório interno do modelo quando comparados a métodos baseados em atribuição. Por exemplo, *rationales* gerados pelo Llama-3 apresentam um *flip rate* de apenas 22.33%. Esses resultados indicam que explicações produzidas por LLMs podem maximizar plausibilidade para avaliadores humanos, mas não refletem fatores responsáveis pela decisão do modelo. Assim, a exigência de *rationales* durante avaliações pode introduzir complexidade discursiva adicional sem garantir aumento correspondente de fidelidade ou confiabilidade. Além disso, oferecer aos avaliadores humanos explicações que não refletem os processos internos dos modelos pode induzi-los ao Viés de Acquiescência ou promover o Efeito de Ancoragem, já discutidos.

4.5. Princípio da Clareza

O Princípio da Clareza estabelece a necessidade de termos **diretrizes claras** sobre a avaliação a ser realizada pelo juiz. Em termos práticos, isso significa (i) ter diretrizes (ii) não ambíguas e (iii) com exemplos e casos de uso.

Prática já estabelecida na linguística de corpus e na anotação de dados em PLN [Freitas 2024], a necessidade de termos instruções claras para a avaliação das saídas das LLMs também já é reconhecida pela indústria[4].

O trabalho de [Pradhan et al. 2021] propõe uma construção iterativa para reduzir ambiguidades nas diretrizes de captura de julgamento, o *framework* FIND-Resolve-Label. Ele consiste em identificar dados ambíguos, resolver os casos e atualizar as diretrizes. A tarefa testada é simples, classificação binária de imagens, realizada com dezenas de anotadores do Mechanical Turk. Apesar da simplicidade da tarefa, a estratégia proposta

[4]`https://www.twine.net/blog/what-is-an-llm-evaluation-rubric/`
`https://www.ai21.com/glossary/foundational-llm/llm-as-a-judge/`

consegue aumentar a acurácia da anotação de aproximadamente 70% para 90%, medida contra um *ground truth* não disponível para os anotadores.

Quanto à captura de avaliações humanas, o trabalho de [Laux et al. 2024] apresenta um experimento com 307 anotadores variando as diretrizes recebidas (regras, regras incompletas e padrões vagos). O estudo também investiga o impacto do retorno financeiro da tarefa e revela que a acurácia dos julgamentos aumentou em 14% quando os humanos têm diretrizes claras em comparação a quando recebem apenas padrões. Em relação ao retorno financeiro, a diferença na acurácia é de apenas 2 pontos. Outro experimento em anotação de inferência em linguagem natural também revela que diretrizes detalhadas, com exemplos e definições técnicas dos fenômenos investigados, melhoram em 17% a anotação de relações semânticas como acarretamento e contradição [Kalouli et al. 2019].

Considerando LLMs, [Agarwal et al. 2024] investiga o papel dos exemplos nos prompts, explorando estratégias de *many-shots* que chegam a ter mais de mil exemplos, testados com o modelo Gemini 1.5 Pro, cuja janela de contexto chega a 1 milhão de tokens. Em todas as 11 tarefas investigadas, incluindo tradução automática e análise de sentimentos, a performance do modelo aumenta quando são oferecidos mais exemplos.

Em resumo, para qualquer tarefa que necessite da captura de julgamento, seja ele humano ou não, é necessário ter diretrizes e quanto mais claras elas forem, menos ruídos ou erros de julgamento estarão presentes nos dados finais.

5. Considerações e Conclusões

Visando criar uma ponte entre áreas que tradicionalmente estudam o conhecimento humano e a literatura sobre avaliação de sistemas baseados em PLN/IA, apresentamos cinco princípios simples, resumidos na Tabela 1. Estes princípios, baseados em conhecimento prévio e consistentes com outras áreas, devem ser integrados à prática de avaliação em PLN/IA. Na Tabela 1, encontram-se também possíveis formas de implementar estes princípios na prática, embora estas implementações recomendadas não sejam exaustivas.

Tabela 1. Resumo dos Princípios Metodológicos para Avaliação de Sistemas Baseados em IA

Princípio	Problema Mitigado	Implementação Recomendada
Independência	Viés de ancoragem, falso consenso, acquiescência e agreement bias	Avaliações às cegas
Diversidade	Viés individual ou cultural; inconsistências de julgamento	Uso de múltiplos avaliadores humanos e/ou múltiplas LLMs-juízes
Confiabilidade	Baixa reprodutibilidade e inconsistências	Mensuração contínua de acordo inter-anotadores (IAA)
Simplicidade	Sobrecarga cognitiva, racionalizações artificiais e baixa consistência	Tarefas objetivas e simples; evitar pedidos desnecessários de explicações
Clareza	Ambiguidade interpretativa e divergência de critérios	Criação de diretrizes explícitas, não ambíguas, com exemplos e casos limítrofes

Considerando que a avaliação em PLN/IA é fundamentalmente um problema de captura de julgamento humano, este trabalho focou em aspectos epistemológicos, metodológicos e operacionais no que toca à metodologia de captura do julgamento, seja este humano ou não. Oferecemos maneiras efetivas de operacionalizar a avaliação. Esperamos que este trabalho seja útil à comunidade, inspirando avaliações mais robustas e fidedignas.

6. Limitações

Neste trabalho, abordamos aspectos operacionais de captura de julgamento para avaliação. Esses aspectos são relevantes tanto em abordagens centradas no humano quanto em abordagens integralmente automatizadas. No entanto, aspectos muito relevantes da avaliação não foram abordados neste texto: taxonomia, esquemas, representatividade e *sampling*, ferramentas de gerenciamento ou versionamento, rastreabilidade, fluxo de trabalho de anotação (iterativo ou não), calibração de avaliadores, casos limítrofes e adversários, governança dos dados, documentação, detecção e mitigação de demais vieses, transparência, responsabilidade e reusabilidade. Também não abordamos questões fundamentais da avaliação humana como a definição do humano-referência (o especialista, o usuário, a população afetada pela aplicação) e as condições de trabalho dos anotadores, como remuneração justa e tempo adequado para o trabalho. Quanto à avaliação automática, aspectos como os impactos ambientais da etapa e a responsabilização pela qualidade da avaliação automática ficaram fora do escopo deste trabalho.

Este trabalho não inclui validação empírica dos princípios propostos, configurando-se antes como um artigo de posicionamento baseado em fundamentos da área de linguística de *corpus* e áreas correlatas que se dedicam à captura da percepção humana e à estruturação dessa percepção em grandes conjuntos de dados *machine-readable*. Embora muito dos princípios encontrem referencial teórico em diferentes áreas do conhecimento, nenhuma dessas áreas foi exaustivamente explorada; conceitos relevantes para as mesmas podem não ter sido investigados integralmente. Trabalhos futuros devem operacionalizar e mensurar o impacto da adoção destes princípios em pipelines reais de avaliação, especialmente na avaliação automatizada. Ressaltamos, porém, que validar cada um destes princípios não será trivial, exigindo experimentos extensos, abrangendo diferentes domínios e tarefas de avaliação, idealmente em diferentes línguas, com a participação de juízes distintos, com base em metodologia rigorosa de comparação.

7. Ética e Uso de Inteligência Artificial

Entendemos que esta pesquisa lança bases para uma metodologia integrada de avaliação de conteúdos gerados por IA, especialmente baseada em conhecimento anterior, provindo de áreas de expertise diferentes da IA em si. Neste contexto, é importante considerar que há inúmeras outras formas de pesquisa que não foram consideradas. Apesar de termos tentado ser abrangentes, é possível, e esperado, que ela não seja exaustiva. Há ainda muitos métodos e muitas áreas a explorar.

Durante esta pesquisa, utilizamos o assistente do Gemini para refinar buscas bibliográficas, o que se mostrou muito útil para conhecer princípios e nos aprofundar em áreas desconhecidas. Utilizamos também ferramentas de IA para validar nossas referências bibliográficas.

O texto deste artigo foi integralmente escrito por seres humanos.

8. Agradecimentos

Trabalho parcialmente apoiado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) Código Financeiro 001; pela Fundação de Amparo à Pesquisa do Estado do Amazonas (FAPEAM) por meio do programa POSGRAD 2025, pelo Projeto NeuralBond (UNIVERSAL 2023, Proc. 01.02.016301.04300/2023-04); pelo CNPq no âmbito do INCT IAIA (406417/2022-9), INCT TILDIAR (408490/2024-1), Projeto AIDEn (Proc. 400936/2025-9) e por bolsa Produtividade em Pesquisa para Altigran da Silva (301925/2025-9).

Referências

Abercrombie, G., Dinkar, T., Curry, A. C., Rieser, V., and Hovy, D. (2025). Consistency is key: Disentangling label variation in natural language processing with intra-annotator agreement.

Agarwal, R., Singh, A., Zhang, L. M., Bohnet, B., Chan, S., Anand, A., Abbas, Z., Nova, A., Co-Reyes, J. D., Chu, E., Behbahani, F., Faust, A., and Larochelle, H. (2024). Many-shot in-context learning. *arXiv preprint arXiv:2404.11018*.

Artemova, E., Tsvigun, A., Schlechtweg, D., Fedorova, N., Tilga, S., and Obmoroshev, B. (2025). Hands-on tutorial: Labeling with llm and human-in-the-loop. In *Proceedings of COLING 2025*.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Asli, C., Clark, E., and Gao, J. (2020). Evaluation of text generation: A survey. *arXiv:2006.14799*.

Bayerl, P. S. and Paul, K. I. (2011). What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics*, 37(4):699–725.

Beloff, H. (1958). Two forms of social conformity: Acquiescence and conventionality. *The Journal of Abnormal and Social Psychology*, 56(1):99–104.

Bird, S. and Liberman, M. (2001). A formal framework for linguistic annotation. *Speech Communication*, 33(1–2):23–60.

Blagec, K., Dorffner, G., Moradi, M., Ott, S., and Samwald, M. (2022). A global analysis of metrics used for measuring performance in natural language processing. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 52–63, Dublin, Ireland. Association for Computational Linguistics.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal. Association for Computational Linguistics.

BRASIL. Conselho Nacional de Justiça (2025). *Resolução n 615, de 10 de fevereiro de 2025*. Conselho Nacional de Justiça, Brasília, DF.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the role of bleu in machine translation research. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Chen, D., Chen, R., Zhang, S., Wang, Y., Liu, Y., Zhou, H., Zhang, Q., Wan, Y., Zhou, P., and Sun, L. (2024). MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.

Chen, O., Castro-Alonso, J., Paas, F., and Sweller, J. (2018). Extending cognitive load theory to incorporate working memory resource depletion: Evidence from the spacing effect. *Educational Psychology Review*, 30(2):483–501.

Choi, J., Hong, Y., and Kim, B. (2025). People will agree what i think: Investigating LLM's false consensus effect. In *Findings of the Association for Computational Linguistics: NAACL 2025*. Association for Computational Linguistics. arXiv:2407.12007v2.

Confident AI (2024). Deepeval: A framework for evaluating large language models. Accessed: 2026-05-12.

de Jong, T. (2010). Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science*, 38(2):105–134.

Durkin, K. (1996). Peer pressure. In Manstead, A. S. R. and Hewstone, M., editors, *The Blackwell Encyclopedia of Social Psychology*. Blackwell Publishers, Oxford.

Fanous, A., Goldberg, J., Agarwal, A. A., Lin, J., Zhou, A., Daneshjou, R., and Koyejo, S. (2025). Syceval: Evaluating llm sycophancy. *arXiv preprint arXiv:2502.08177*.

Fayyaz, M., Yin, F., Sun, J., and Peng, N. (2024). Evaluating human alignment and model faithfulness of llm rationale.

Foody, G. M. (2024). Ground truth in classification accuracy assessment: Myth and reality. *Geomatics*, 4(1):81–90.

Fort, K., Adda, G., and Cohen, K. B. (2011). Last words: Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420.

Freitas, C. (2024). Dataset e corpus. In *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, chapter 16. BPLN, 4 edition.

Frenda, S., Abercrombie, G., Basile, V., Pedrani, A., Panizzon, R., Cignarella, A. T., Marco, C., and Bernardi, D. (2024). Perspectivist approaches to natural language processing: a survey: Perspectivist approaches to natural language processing... *Lang. Resour. Eval.*, 59(2):1719–1746.

Gao, M., Hu, X., Yin, X., Ruan, J., Pu, X., and Wan, X. (2025). LLM-based NLG evaluation: Current status and challenges. *Computational Linguistics*, 51:661–687.

Gilbert, D. T. (1991). How mental systems believe. *American Psychologist*, 46(2):107–119.

Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., and Kamar, E. (2022). Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection.

Jacovi, A. and Goldberg, Y. (2020). Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4212, Online. Association for Computational Linguistics.

Jain, S., Ahmed, U. Z., Sahai, S., and Leong, B. (2025). Beyond consensus: Mitigating the agreeableness bias in llm judge evaluations.

Jang, M. E. and Silavong, F. (2025). InstaJudge: Aligning judgment bias of LLM-as-judge with humans in industry applications. In Potdar, S., Rojas-Barahona, L., and Montella, S., editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1158–1172, Suzhou (China). Association for Computational Linguistics.

Kalouli, A.-L., Buis, A., Real, L., Palmer, M., and de Paiva, V. (2019). Explaining simple natural language inference. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 132–143, Florence, Italy. Association for Computational Linguistics.

Kalouli, A.-L., Real, L., and de Paiva, V. (2017). Textual inference: getting logic from humans. In Gardent, C. and Retoré, C., editors, *Proceedings of the 12th International Conference on Computational Semantics (IWCS) — Short Papers*, Montpellier, France. Association for Computational Linguistics.

Laux, J., Stephany, F., and Liefgreen, A. (2024). Improving task instructions for data annotators: How clear rules and higher pay increase performance in data annotation in the ai economy.

Lee, Y., Kim, J., Kim, J., Cho, H., Kang, J., Kang, P., and Kim, N. (2025). Checkeval: A reliable llm-as-a-judge framework for evaluating text generation using checklists. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Suzhou, China. Association for Computational Linguistics.

Li, K., Li, Y., Zhang, T., Luo, H., Wu, X., Glass, J. R., and Meng, H. M. (2025a). Ragzeval: Enhancing rag response evaluation through end-to-end reasoning and ranking-based reinforcement learning. In *Proceedings of EMNLP 2025*.

Li, S., Li, J., Qu, Y., Shi, X., Guo, Y., He, Z., Wang, Y., and Tan, W. (2025b). Semantic-eval : A semantic comprehension evaluation framework for large language models generation without training. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9675–9690, Vienna, Austria. Association for Computational Linguistics.

Lin, S., Hilton, J., and Evans, O. (2021). Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Madsen, A. (2024). New faithfulness-centric interpretability paradigms for natural language processing. *arXiv preprint arXiv:2411.17992*.

Marasović, A. (2018). Nlp's generalization problem, and how researchers are tackling it. *The Gradient*. Accessed: 2026-05-14.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., and Zamparelli, R. (2014). A SICK cure for the evaluation of compositional distributional semantic mo-

dels. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Marinho, M., Vianna, D., Migliorini, G., Real, L., Ramalhete, M., and da Silva, A. (2026). Grounding the evaluation of ai-generated legal answers: The hermeval protocol and empirical evidence. In *Proceedings of the International Conference on Artificial Intelligence and Law (ICAIL)*, Singapore. ACM.

Martin, D., Hanrahan, B. V., O'Neill, J., and Gupta, N. (2014). Being a turker. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, page 224–235, New York, NY, USA. Association for Computing Machinery.

McEnery, T. and Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press, Cambridge, 2 edition.

Messick, S. and Jackson, D. N. (1961). Acquiescence and the factorial interpretation of personality tests. In Jackson, D. N. and Messick, S., editors, *Psychological Measurement in Personality Research*, pages 61–103. University of Chicago Press, Chicago.

Meyer, C. F. (2002). *English Corpus Linguistics: An Introduction*. Cambridge University Press, Cambridge.

Novikova, J., Dušek, O., Cercas Curry, A., and Rieser, V. (2017). Why we need new evaluation metrics for NLG. In Palmer, M., Hwa, R., and Riedel, S., editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Peixoto, R. M., Guimarães, G. A. S., Cruz, G. O. R., Araujo, M. M., Santos, L. L., Cruz, M. A. S., Oliveira, E. L. S., Winkler, I., and Nascimento, E. G. S. (2023). Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data and Cognitive Computing*, 7(1).

Pavlick, E. and Kwiatkowski, T. (2019). Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Plank, B. (2022). The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pradhan, V. K., Schaekermann, M., and Lease, M. (2021). In search of ambiguity: A three-stage workflow design to clarify annotation guidelines for crowd workers.

Qian, M., Sun, G., Gales, M. J. F., and Knill, K. M. (2026). Who can we trust? llm-as-a-jury for comparative assessment.

Raina, V., Liusie, A., and Gales, M. (2024). Is llm-as-a-judge robust? investigating universal adversarial attacks on zero-shot llm assessment. *arXiv preprint arXiv:2402.14016*.

Reiter, E. and Belz, A. (2009). An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Ross, L., Greene, D., and House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, 13(3):279–301.

Sandan, I. B., Dinh, T. A., and Niehues, J. (2025). Knockout LLM assessment: Using large language models for evaluations through iterative pairwise comparisons. In Arviv, O., Clinciu, M., Dhole, K., Dror, R., Gehrmann, S., Habba, E., Itzhak, I., Mille, S., Perlitz, Y., Santus, E., Sedoc, J., Shmueli Scheuer, M., Stanovsky, G., and Taffjord, O., editors, *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 121–128, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP 2008*, Honolulu, Hawaii. Association for Computational Linguistics.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285.

Thakur, A. S., Choudhary, K., Ramayapally, V. S., Vaidyanathan, S., and Hupkes, D. (2024). Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *arXiv preprint arXiv:2406.12624*.

Trautmann, D., Ostapuk, N., Grail, Q., Pol, A. A., Bonifazi, G., Gao, S., and Gajek, M. (2024). Measuring the groundedness of legal question-answering systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

Verga, P., Hofstätter, S., Althammer, S., Su, Y., Piktus, A., Arkhangorodsky, A., Xu, M., White, N., and Lewis, P. (2024). Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*.

Voutilainen, A. (2012). Improving corpus annotation productivity: a method and experiment with interactive tagging. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2097–2102, Istanbul, Turkey. European Language Resources Association (ELRA).

Wei, H., He, S., Xia, T., Liu, F., Wong, A., Lin, J., and Han, M. (2025). Systematic evaluation of llm-as-a-judge in llm alignment tasks: Explainable metrics and diverse prompt templates.

Wojcieszak, M. and Price, V. (2009). What underlies the false consensus effect? how personal opinion and disagreement affect perception of public opinion. *International Journal of Public Opinion Research*, 21(1):25–46.

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., and He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381.

Yang, J., Redi, J., Demartini, G., and Bozzon, A. (2016). Modeling task complexity in crowdsourcing. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 4(1):249–258.

Zhao, W. X., Zhou, K., Li, J., Tang, T., and Wen, J.-R. (2026). *Human Alignment*, pages 209–256. Springer Nature Singapore, Singapore.

Zheng, L., Chiang, W.-L., Sheng, Y., and et al. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS Workshop on Instruction Tuning and Instruction Following*.

Zhu, Z., Tieleman, O., Bukhtiyarov, A., and Chen, J. (2026). Cyclicjudge: Mitigating judge bias efficiently in llm-based evaluation.