

**Avaliação do Conhecimento dos Grandes Modelos de
Linguagem Públicos sobre a Laminação Controlada
de Aços Microligados**

Antonio Augusto Gorni

Trabalho de Conclusão de Curso
MBA em Inteligência Artificial e Big Data

UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Avaliação do Conhecimento dos
Grandes Modelos de Linguagem
Públicos sobre a Laminação
Controlada de Aços Microligados

Antonio Augusto Gorni

Antonio Augusto Gorni

Avaliação do Conhecimento dos Grandes Modelos de Linguagem Públicos sobre a Laminação Controlada de Aços Microligados

Trabalho de conclusão de curso apresentado ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial.

Orientador: Fernando Pereira dos Santos

USP - São Carlos

2026

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

G671a Gorni, Antonio Augusto
 Avaliação do Conhecimento dos Grandes Modelos de
 Linguagem Públicos sobre a Laminação Controlada de
 Aços Microligados / Antonio Augusto Gorni;
 orientador Fernando Pereira dos Santos. -- São
 Carlos, 2026.
 61 p.

 Trabalho de conclusão de curso (MBA em
 Inteligência Artificial e Big Data) -- Instituto de
 Ciências Matemáticas e de Computação, Universidade
 de São Paulo, 2026.

 1. Inteligência Artificial. 2. Grandes Modelos de
 Linguagem. 3. Aços Microligados. 4. Laminação
 Controlada. I. dos Santos, Fernando Pereira ,
 orient. II. Título.

DEDICATÓRIA

*À memória de minha avó paterna,
Emma Falco, que, com sua visão e
determinação, logrou arrancar a
família da lavoura arcaica e inseri-
la no Século XX.*

AGRADECIMENTOS

À minha esposa **Regina**, pelo amor, apoio, incentivo e paciência incondicionais, não só durante o desenvolvimento deste trabalho, como também ao longo de toda a minha carreira profissional e de nossa vida em comum.

Ao Prof. Dr. **Fernando Pereira dos Santos**, pela orientação sempre segura, e a atenção e apoio contínuos.

À Profa. Dra. **Solange Oliveira Resende**, pelo suporte constante e entusiástico durante o Curso, orientando seu andamento e reforçando a minha motivação nos momentos menos inspirados.

Aos **professores e tutores** do MBA em Inteligência Artificial e Big Data, pelo compartilhamento de seu conhecimento, resolução das dúvidas e “insights” sobre todos os temas envolvidos, bem como ao **pessoal do suporte acadêmico**, sempre prestativo, cortês e eficiente na solução das dificuldades logísticas do curso.

Aos **colegas da quinta turma** do MBA em Inteligência Artificial e Big Data, pelo contínuo suporte técnico e humano ao longo de todo o curso – e além. Apesar de, infelizmente, não termos nos conhecido pessoalmente, a comunidade on-line formada no grupo de What’sApp sempre esteve muito atuante no auxílio mútuo.

Finalmente, ao **Instituto de Ciências Matemáticas e de Computação (ICMC)** da USP, em São Carlos SP, pela sua longa tradição no oferecimento de cursos de extensão à comunidade. Em janeiro de 1977, decidido a reforçar minha incipiente formação na promissora área de computação, fiz o curso de extensão “Linguagens de Programação”, oferecido pelo seu antecessor, o Instituto de Ciências Matemáticas de São Carlos (ICMSC). Perdi minhas férias de verão, mas consegui obter uma sólida formação na área de programação, a qual me foi extremamente útil ao longo de minha carreira profissional. O ciclo se repete agora, meio século depois, com o MBA em Inteligência Artificial e Big Data, cujos temas são ainda mais revolucionários e disruptores. Desta vez eu nem precisei me deslocar para São Carlos para participar do curso, que foi totalmente *on-line*. Finalmente, vale lembrar que a elaboração de programas digitais está sendo assumida cada vez mais pelos próprios computadores, o que ameaça tornar obsoleto meu curso de 1977. *Brave new world!*

EPÍGRAFE

Danger!

Let the bridge computer speak...

Stranger!

Load your program. I am yourself.

No computer will stand in my way,

Only blood can cancel my pain,

Guardians of a new clear dawn!

Let the maps of war be drawn.

(...)

Rejoice, glory is ours!

Our young men have not died in vain,

Their graves need no flowers,

The tapes have recorded their names.

I am all there is...

Negative! Primitive! Limited! I... let you live!

But I gave you life!

What else could you do?

To do what was right!

I'm perfect! Are you?

Greg Lake & Peter Sinfield: Karn Evil 9, Third Impression (1973)

RESUMO

GORNI, A. A. **Avaliação do Conhecimento dos Grandes Modelos de Linguagem Públicos sobre a Laminação Controlada de Aços Microligados.** 2026. 61 f. Monografia (MBA em Ciências de Dados) – Centro de Ciências Matemáticas Aplicadas à Indústria, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2026.

Nos últimos anos as aplicações de inteligência artificial disseminaram-se explosivamente, especialmente os *chatbots* baseados em grandes modelos de linguagem (*Large Language Models*, LLMs), disponibilizados de forma pública e com versões gratuitas. Contudo, esses recursos tendem a sofrer as chamadas “alucinações”, ou seja, elaboram respostas envolventes e muito bem redigidas, mas totalmente fora da realidade, por exemplo, quando inquiridos sobre temas que não foram incluídos em seu aprendizado. A técnica mais imediata e de custo virtualmente zero para se evitar esse problema é a engenharia de *prompts*, que foi aplicada neste trabalho, o qual teve por objetivo verificar o conhecimento dos atuais *chatbots* sobre a laminação controlada de aços estruturais microligados. Foram testados cinco desses assistentes conversacionais populares baseados nos grandes modelos de linguagem, em sua versão gratuita: ChatGPT, Co-Pilot, Gemini, Claude e DeepSeek, os quais foram submetidos a um questionário sobre esse tema. Uma vez que a avaliação humana das respostas desse questionário mostrou-se inviável na prática, adotou-se a abordagem *LLM-as-a-Judge* (LLM como Juiz), onde as respostas de um *chatbot* ao questionário eram corrigidas pelos quatro restantes. Os resultados obtidos foram razoáveis em termos de qualidade, com ampla variedade de nível de detalhamento entre os modelos aqui testados, com destaque positivo para o Claude e Gemini. Os resultados obtidos permitem concluir que se aplica aos resultados dos *chatbots* a mesma leitura crítica que se deve ter ao ler artigos técnicos escritos por humanos, ainda que seja recomendável uma curadoria por parte de especialistas experientes nessa área, já que foi inevitável a ocorrência de algumas alucinações dentro do procedimento automatizado e massivo aqui adotado para se fazer a avaliação dos modelos. Portanto, ao se consultar um *chatbot*, não se deve aceitar incondicionalmente a primeira resposta obtida. A inteligência artificial deve ser um interlocutor intelectual, e não uma pitonisa. O humano precisa estar no laço, elaborando múltiplos *prompts* interativos ou estabelecendo fluxos organizados de pensamento.

Palavras-chave: inteligência artificial; grandes modelos de linguagem; aços microligados; laminação controlada.

ABSTRACT

GORNI, A.A. Assessment of Knowledge of Large Public Language Models on Controlled Rolling of Microalloyed Steels. 2026. 61 f. Monografia (MBA em Ciências de Dados) – Centro de Ciências Matemáticas Aplicadas à Indústria, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, Brazil, 2026.

In recent years, artificial intelligence applications have proliferated explosively, especially chatbots based on Large Language Models (LLMs), which are publicly available and offered in free versions. However, these applications tend to suffer from so-called "hallucinations," i.e., they produce engaging and well-written responses, but completely out of touch with reality, for example, when questioned about topics not included in their learning. The most immediate and virtually cost-free technique to avoid this problem is prompt engineering, which was applied in this study aimed at verifying the knowledge of current LLMs about the controlled rolling of microalloyed structural steels. Five of the most common chatbots were evaluated in their free versions: ChatGPT, CoPilot, Gemini, Claude, and DeepSeek, which were subjected to a questionnaire on this topic. Since human evaluation of the questionnaire responses proved impractical, the LLM-as-a-Judge approach was adopted, where the responses of one chatbot were corrected by the remaining four. The results obtained were reasonable in terms of quality, with a wide variety of detailed levels among the LLMs assessed, with positive highlights on Claude and Gemini. The results allow us to conclude that the same critical reading that should be applied to technical articles written by humans applies to chatbot results, although specialized human curation is recommended if they are used professionally, given that some hallucinations were inevitable within the automated and massive procedure adopted here for evaluating the models. So, when consulting a chatbot, one should not unconditionally accept the first response obtained. Artificial intelligence should be an intellectual interlocutor, not a fortune teller. The human must be in the loop, developing multiple interactive or chain-of-thought prompts.

Keywords: artificial intelligence; large language models; microalloyed steels; controlled rolling.

LISTA DE ILUSTRAÇÕES

Figura 1 – Laminador de Chapas Grossas em operação.....	23
Figura 2 – Evolução aproximada, ao longo do tempo, das quatro gerações de modelos de linguagem, do ponto de vista da sua capacidade de resolução de tarefas.....	28
Figura 3 – Lançamento de grandes modelos de linguagem representativos ao longo do tempo. Os modelos destacados em amarelo disponibilizam publicamente os parâmetros do modelo treinado.....	28

LISTA DE TABELAS

Tabela 1 – Comparação entre as abordagens de “Fine Tuning”, RAG e a Engenharia de <i>Prompts</i>	32
Tabela 2 – Número de caracteres (exceto espaços) das respostas às questões geradas pelos LLMs candidatos. DPR significa Desvio Padrão Relativo em relação à média, expresso em percentagem.....	41
Tabela 3 – Número total de caracteres (exceto espaços), tempo dispendido, velocidade e número de páginas ocupadas pelas respostas às questões geradas pelos LLMs candidatos.....	41
Tabela 4 – Número total de caracteres (exceto espaços), das correções às respostas ao questionário geradas pelos LLMs avaliadores.....	42
Tabela 5 – Número total de caracteres (exceto espaços), tempo dispendido, velocidade e número de páginas ocupadas pelas correções às respostas das perguntas geradas pelos LLMs avaliadores.....	43
Tabela 6 – Notas concedidas pelos LLMs avaliadores para as respostas ao questionário que foram fornecidas pelos LLMs candidatos, de acordo com os itens de avaliação considerados, e sua média global.....	44
Tabela 7 – Notas globais concedidas pelos LLMs avaliadores aos LLM candidatos.....	45
Tabela 8 – Veredictos recebidos pelos LLMs candidatos sobre sua contratação como supervisor técnico de uma instalação de laminação a quente.....	46

LISTA DE ABREVIATURAS E SIGLAS

AI	–	Artificial Intelligence
API	–	Application Programming Interface
BERT	–	Bidirectional Encoder Representations from Transformers
DPR	–	Desvio Padrão Relativo (em relação à média)
ELMo	–	Embeddings from Language Models
GPT	–	Generative Pre-Trained Transformer
GPU	–	Graphics Processing Unit
HTML	–	HyperText Markup Language
IA	–	Inteligência Artificial
IDE	–	Integrated Development Environment
LLM	–	Large Language Models
LM	–	Language Modeling
LSTM	–	Long-Short Term Memory
NLM	–	Neural Language Models
NLP	–	Natural Language Processing
PLM	–	Pre-Trained Language Models
RAG	–	Retrieval Augmented Generation
RLHF	–	Reinforcement Learning from Human Feedback
RLT	–	Recrystallization Limit Temperature
RST	–	Recrystallization Stop Temperature
SLM	–	Statistical Language Models
TMCP	–	Thermomechanical Controlled Processing
T _{nr}	–	Temperature of No-Recrystallization
XML	–	Extensible Markup Language

LISTA DE SÍMBOLOS

°C Graus Celsius

SUMÁRIO

1 INTRODUÇÃO.....	21
1.1 A Motivação para este Trabalho.....	21
1.2 Objetivos deste Trabalho.....	22
1.3 Artigo Baseado neste Trabalho Aceito para Apresentação em Congresso.....	22
2 FUNDAMENTAÇÃO TEÓRICA.....	23
2.1 O Contexto Metalúrgico.....	23
2.2 A Origem e Evolução dos Grandes Modelos de Linguagem.....	26
2.2.1 ChatGPT.....	29
2.2.2 CoPilot.....	30
2.2.3 Gemini.....	30
2.2.4 Claude.....	30
2.2.5 DeepSeek.....	31
2.3 Aplicação dos Grandes Modelos de Linguagem na Siderurgia.....	31
2.4 Metodologias para Avaliação do Desempenho de LLMs.....	34
3 METODOLOGIA.....	37
3.1 Premissas Básicas.....	37
3.2 Procedimento.....	37
3.3 Disponibilização dos Dados Usados Neste Trabalho.....	39
4 ANÁLISE E RESULTADOS.....	40
4.1 Respostas dos LLMs aos Questionários.....	40
4.2 Correções das Respostas ao Questionário Geradas pelos LLMs Candidatos.....	41
4.3 Avaliação das Respostas ao Questionário.....	43
4.4 Qualidade das Respostas e Avaliações do Ponto de Vista Metalúrgico.....	47
5 CONCLUSÃO.....	53
REFERÊNCIAS.....	55
APÊNDICE A – Questionário Apresentado aos LLMs.....	58
APÊNDICE B – Procedimento para Avaliação das Respostas dos LLMs ao Questionário.....	60
APÊNDICE C – Declaração sobre o Uso de Inteligência Artificial neste Trabalho.....	61

1 INTRODUÇÃO

1.1 A Motivação para este Trabalho

Aços planos estruturais laminados a quente, com níveis altos e precisos de resistência mecânica, vem encontrando demanda cada vez maior, uma vez que eles permitem a chamada “desmaterialização” das estruturas feitas com eles. Seu maior nível de resistência mecânica permite reduzir a seção transversal dos componentes – e, conseqüentemente, seu peso. Dessa forma, estruturas estáticas irão requerer menor esforço para serem erigidas; caso o aço seja usado em veículos, seu menor peso reduzirá seu consumo de combustível e, conseqüentemente, suas emissões de gás carbônico. Tudo isso está impulsionando o interesse por soluções baseadas em inteligência artificial que otimizem a concepção, custo e processamento desse tipo de produto.

A justificativa para este trabalho baseia-se na relevância econômica e tecnológica desse setor. Apenas no Brasil, a produção de aços laminados a quente foi da ordem de 23,8 milhões de toneladas em 2024, segundo dados do Anuário Estatístico 2025 (Instituto Aço Brasil, 2025), evidenciando a magnitude do impacto potencial de um sistema inteligente aplicado a essa cadeia produtiva. Os projetos de liga de aços estruturais de alta resistência, por sua vez, representam um domínio técnico complexo, no qual pequenas variações na composição ou nos parâmetros termomecânicos podem alterar significativamente as propriedades do produto final. Dessa forma, ferramentas de suporte à decisão que utilizem informações confiáveis e atualizadas têm potencial para aumentar a competitividade industrial, reduzir custos e a pegada de carbono do produto, e minimizar retrabalho, aspectos fundamentais quando se considera a intensa competição existente no mercado brasileiro por parte dos similares importados, especialmente da China.

Nesse contexto, é muito atraente a proposta do desenvolvimento de modelos especialistas, confiáveis e consistentes, baseados em técnicas de inteligência artificial, que elaborem recomendações técnicas no processo de fabricação de aços planos estruturais com alta resistência feitos com aços microligados. Por outro lado, é necessário considerar que essa abordagem requereria muito esforço e tempo para ser implementada, envolvendo equipes altamente especializadas, além de necessitar de servidores com alta capacidade para funcionar adequadamente e atualização constante para que mantenha informações atualizadas sobre o tema. Tudo isso também implicaria em altos custos de implantação e operação. Portanto, é necessário que haja uma

justificativa muito bem fundamentada para a concretização dessa ferramenta de inteligência artificial.

Dada essa situação, antes de partir para esta situação complexa, seria muito oportuno averiguar se não haveria alternativas de execução mais rápida e com menor custo para se conseguir obter esse conhecimento, utilizando as informações disponíveis que foram utilizadas no treinamento dos grandes modelos de linguagem públicos.

1.2 Objetivos deste Trabalho

Avaliar o desempenho dos atuais assistentes conversacionais (*chatbots*) de inteligência artificial, com acesso público e gratuito, baseados em grandes modelos de linguagem, na área do conhecimento sobre o processo de fabricação de aços planos estruturais com alta resistência feitos com micro-ligas de nióbio, titânio e vanádio, visando determinar quais deles são os melhores nesse sentido. A partir dos resultados obtidos pode-se identificar melhor quais as áreas poderiam ser atendidas por modelos específicos de inteligência artificial, concentrando-se os esforços necessários em soluções mais complexas onde houver maior carência de conhecimento por parte dos grandes modelos de linguagem comerciais.

A ideia de utilizar as versões gratuitas dos *chatbots* tem como objetivo tornar este trabalho útil ao maior número de profissionais envolvidos nessa área, em particular os que possuem menor poder aquisitivo, como estudantes, técnicos e engenheiros em início de carreira, requerendo apenas acesso à internet e algum conhecimento sobre os *prompts* mais eficazes para lidar com os *chatbots*, de forma a extrair conhecimento genuíno e minimizar suas alucinações típicas.

1.3 Artigo Baseado neste Trabalho Aceito para Apresentação em Congresso

Um artigo baseado nos resultados e conclusões aqui descritos foi aceito para apresentação no 61º Seminário de Laminação, Conformação de Metais e Produtos - parte da ABM Week 10 (<https://abmbrasil.com.br/evento/abm-week-10>), evento promovido pela Associação Brasileira de Metalurgia, Materiais e Mineração, a ser realizado em São Paulo SP entre os dias 8 a 10 de setembro de 2026, sob o título *Assessment of the Knowledge of Public Large Language Models About Thermomechanical Controlled Processing of Microalloyed Steels*.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 O Contexto Metalúrgico

As chapas estruturais de aço são feitas através do processo de laminação a quente, onde um semi-produto siderúrgico chamado placa, com espessura entre 180 e 250 mm, largura entre 1.500 e 2.000 mm e comprimento entre 3.000 e 12.000 mm, é aquecido até temperaturas da ordem de 1100-1200°C e, a seguir, tem sua espessura reduzida através de passagens sucessivas por laminadores, até atingir os valores finais requeridos para o produto final que, na maior parte dos casos, se situa entre 2 e 50 mm. A temperatura final do processo varia entre 800 e 900°C. A Figura 1 mostra um Laminador de Chapas Grossas em operação (Bolota, 2024).

Figura 1: Laminador de Chapas Grossas em operação.



(Fonte: Bolota, 2024)

Até a década de 1950 as chapas estruturais laminadas a quente eram feitas com aços contendo teores relativamente baixos de carbono, da ordem de 0,10 a 0,30%, mais outros elementos de liga como manganês, silício e cromo, entre outros. Nessa época o papel da laminação era exclusivamente de conformação mecânica, ou seja, alcançar as espessuras finais desejadas para o produto. Não havia um controle refinado das condições de processo, buscando-se tão somente viabilizá-lo em função da capacidade disponível no equipamento e da composição química do aço que estava sendo produzido. Nos casos em que os produtos laminados precisassem atender a especificações de propriedades mecânicas estabelecidas por normas técnicas e/ou requeridas pelos usuários, como resistência mecânica e tenacidade, era necessário submetê-los a um tratamento térmico adicional, de normalização ou têmpera mais revenido. Esse tratamento adicional implica em aumento de trabalho, custos e prazo para entrega do produto.

Na década de 1960 surgiram os chamados aços microligados, são produtos siderúrgicos onde os elementos de liga presentes em sua composição são substituídos parcialmente por outros elementos como nióbio, titânio e vanádio, presentes em teores da ordem de 0,010 a 0,050%. Esse tipo de aço permitiu dispensar os tratamentos térmicos posteriores à laminação a quente, desde que os parâmetros desse processo fossem rigorosamente controlados – ou seja, as sucessivas reduções de espessura da placa são aplicadas de forma sincronizada com a evolução decrescente da sua temperatura, visando aproveitar os fenômenos metalúrgicos promovidos pelos elementos de micro-liga. Surgiu assim o chamado tratamento termomecânico controlado (*Thermomechanical Controlled Processing*, TMCP) ou, como é comumente designado, laminação controlada.

Dessa forma, graças ao efeito metalúrgico proporcionado principalmente pelo nióbio, tornou-se possível obter aços com microestrutura mais refinada e com as propriedades mecânicas desejadas diretamente do calor da laminação. Além disso, foi possível reduzir substancialmente o teor de carbono desses aços, o que melhorou bastante a sua tenacidade (ou seja, resistência ao impacto) e soldabilidade.

Mas tudo isso tem um custo: é muito trabalhoso definir a composição química do aço e os melhores parâmetros do processo de laminação a quente de forma a promover, de forma otimizada, os mecanismos metalúrgicos necessários para se conseguir uma microestrutura com máximo grau de uniformidade e refino.

Entre as principais variáveis de processo de laminação controlada dos aços microligados que precisam ser consideradas se incluem:

- A temperatura e tempo de reaquecimento da placa, semi-produto fornecido pela Aciaria, onde deve ocorrer grau adequado de solubilização do nióbio presente no aço e redução da resistência à deformação a quente do material, minimizando os esforços mecânicos decorrentes da laminação;
- As condições da primeira fase do processo, denominada esboçamento ou desbaste, onde se inicia o processo de laminação da placa (que, a partir de agora, passa a ser chamada de esboço de chapa ou apenas esboço), o qual é executado sob temperaturas relativamente altas, acima da chamada RLT (*Recrystallization Limit Temperature*), onde a austenita sofre recristalização plena entre os passes de laminação, mantendo homogênea a microestrutura do aço;
- O período de espera, sob temperaturas mais baixas, onde a laminação tem de ser interrompida pois, se continuasse, a recristalização da austenita entre os passes de laminação seria parcial devido à precipitação atrasada de compostos de nióbio, prejudicando a homogeneidade da microestrutura do material e, conseqüentemente, as propriedades do produto final. Essa etapa de resfriamento deve ocorrer até que se atinja a chamada temperatura *Recrystallization Stop Temperature* (RST), em que a precipitação de compostos de nióbio se torna rápida o suficiente para suprimir totalmente a recristalização da austenita entre os passes de laminação, novamente permitindo a obtenção de uma microestrutura não só muito homogênea, como também muito refinada, contribuindo para se alcançar os níveis de resistência mecânica e de tenacidade requeridos no produto final;
- Nesse momento se inicia a última fase de laminação, denominada acabamento, onde o esboço é laminado até se alcançar a espessura desejada pelo cliente. Devido às relativamente baixas temperaturas em que ela ocorre, o endurecimento por encruamento do esboço é progressivo, elevando significativamente os esforços mecânicos requeridos pelo laminador;
- Após a laminação, o esboço de chapa pode ser resfriado ao ar ou em água até a temperatura ambiente, conforme a composição química do aço e os requisitos mecânicos especificados pelo cliente.

Centenas de trabalhos já foram e muitos outros ainda serão publicados sobre esse assunto, já que técnicas e produtos siderúrgicos inéditos surgem ano após ano para atender às necessidades sempre crescentes de infraestrutura e, mais recentemente, da minimização da pegada de carbono dos produtos siderúrgicos (Villalobos et al., 2018; Hamryszczak, 2024; Mohrbacher 2026).

2.2 A Origem e Evolução dos Grandes Modelos de Linguagem

O modelamento da linguagem (*Language Modelling*, LM) constitui uma das principais abordagens para implantar a inteligência de linguagem em máquinas. Ele tem como objetivo geral modelar a semelhança generativa das sequências de palavras, de forma a prever as probabilidades de palavras futuras ou faltantes. Serão descritos a seguir a evolução das quatro etapas do desenvolvimento dos modelos de linguagem ao longo do tempo (Zhao et al., 2026).

Os **Modelos Estatísticos de Linguagem** (*Statistical Language Models*, SLM) foram desenvolvidos com base nos métodos estatísticos de aprendizado que surgiram na década de 1990. A ideia básica por trás deles consiste na construção de um modelo para previsão de palavra baseado na premissa de Markov – por exemplo, prever a próxima palavra com base no contexto mais recente. Esse tipo de modelo tem sido muito aplicado para melhorar as tarefas de recuperação de informação e processamento em linguagem natural. Por outro lado, essa abordagem sofre da chamada maldição da multidimensionalidade: é difícil estimar precisamente modelos de linguagem de alta ordem devido à necessidade de se estimar um número exponencial de probabilidades (Zhao et al., 2026).

Já os **Modelos Neurais de Linguagem** (*Neural Language Models*, NLM) determinam a probabilidade de sequências de palavras usando redes neurais, como o perceptron com múltiplas camadas e redes neurais recorrentes. Foi então proposto o modelo word2vec, constituído por uma rede neural simplificada rasa, para o aprendizado de representações distribuídas de palavras. Este desenvolvimento iniciou o uso de modelos de linguagem para representação de aprendizado e constituiu um grande progresso no campo do processamento de linguagem natural (Zhao et al., 2026).

Os **Modelos de Linguagem Pré-Treinados** (*Pre-Trained Language Models*, PLM) foi a etapa seguinte da evolução, com o desenvolvimento do ELMo (*Embeddings from Language Models* ou vetores de representação provenientes de modelos de linguagem) pelo Allen Institute for AI, que se caracteriza por ser constituído por redes neurais recorrentes bidirecionais com memória de curto e longo prazo (*Long-Short Term Memory*, LSTM). Dessa forma foi resolvido um problema inerente ao Word2Vec, que era a atribuição de um único vetor fixo para cada palavra, criando ambiguidades semânticas. A seguir, por meio da adoção da arquitetura *Transformer* de rede neural, altamente paralelizável e dispondo de mecanismos de auto-atenção, foi proposto o algoritmo BERT (*Bidirectional Encoder Representations from Transformers* ou representações de codificador bidirecional baseadas em *Transformers*) através

do pré-treinamento bidirecional especialmente concebido para modelos de linguagem usando grandes coleções de texto bruto (*corpora*) sem anotações. Essas representações pré-treinadas de palavras sensíveis ao contexto são muito eficazes como características semânticas de uso geral, tendo elevado significativamente o nível de desempenho das tarefas de processamento de linguagem natural (Zhao et al., 2026).

Finalmente, chegou-se nos **Grandes Modelos de Linguagem** (*Large Language Models* ou LLMs), também conhecidos como Modelos Fundacionais (*Foundation Models*), quando pesquisadores verificaram que a expansão do tamanho do modelo ou do conjunto de dados em modelos de linguagem pré-treinados frequentemente aumentava a capacidade do modelo em resolver tarefas. Grandes expansões proporcionavam capacidades surpreendentes na resolução de várias tarefas complexas. Uma aplicação notável desses modelos ocorreu no chamado ChatGPT2, que adaptou os modelos da série GPT para uso em assistentes conversacionais (*chatbots*) e apresentam uma espantosa capacidade de diálogo com seres humanos (Zhao et al., 2026).

Como está mostrado na Figura 2, as tarefas feitas pelos modelos de linguagem iniciais expandiram-se grandemente ao longo do tempo. No início suas atividades estavam associadas a discursos e recuperação de informações, onde as probabilidades previstas pelo modelo podiam melhorar o desempenho dessas tarefas. O uso de redes neurais melhorou o desempenho desses modelos. A seguir, os modelos de linguagem pré-treinados aprenderam representações sensíveis ao contexto, que podem ser otimizadas conforme as tarefas específicas que eles devem executar. Finalmente, os grandes modelos de linguagem podem ser aprimorados pela exploração do efeito de escala na capacidade do modelo, aumentando a sua capacidade a ponto tal que hoje podem ser considerados como executores de tarefas de uso geral. Isso revolucionou a forma como as pessoas usam e desenvolvem modelos de inteligência artificial e levou ao desenvolvimento e eventual disponibilização de inúmeros grandes modelos de linguagem, conforme mostra a Figura 3. A disponibilidade gratuita de vários assistentes conversacionais baseados em inteligência artificial – como, por exemplo, Chat GPT (lançado em 2022), CoPilot (2023), Gemini (2023), Claude (2023) e DeepSeek (2025), constituem um ameaçador competidor aos motores de busca tradicionais – que, por sinal, também estão adotando recursos baseados em inteligência artificial (Zhao et al., 2026).

Apesar desse progresso vertiginoso, o desenvolvimento adicional dessas ferramentas de inteligência artificial precisa vencer alguns obstáculos. Entre eles, o fato de que os princípios que embasam o funcionamento dos grandes modelos de linguagem ainda não foram devidamente explorados e entendidos. Ainda é um mistério a origem das grandes habilidades que eles

determinísticas, ou seja, como são modelos estatísticos, as respostas para uma mesma pergunta serão diferentes entre si ao longo do tempo. É algo parecido com o gato de Schrödinger, ou seja, é necessário realizar o experimento para se saber o resultado final, que não necessariamente será o mesmo ao ser repetido.

Também se deve ter em mente que os LLMs não raciocinam, mas sim, como modelos de linguagem sofisticados, geram texto plausível, mas não necessariamente verdadeiro, dentro do contexto da consulta que está sendo feita (Zhao et al., 2026). Outro problema é a tendência de as LLMs apresentarem sicofancia, ou seja, buscarem a aprovação pelo usuário. Por sua natureza intrínseca, o LLM evita discordar do usuário, mesmo dispondo da informação correta, optando por sacrificar a veracidade de sua resposta. Isso decorre pelo fato do treinamento dos LLMs ter sido geralmente feito por aprendizado por reforço com realimentação humana (*Reinforcement Learning from Human Feedback*, RLHF), onde a concordância do humano com a resposta do LLM é considerada por este como sendo uma recompensa (Ouyang et al, 2022). Tais situações levam à ocorrência das chamadas “alucinações”, ou seja, respostas que soam plausíveis e muito bem articuladas, mas que na verdade são totalmente incorretas. Isso também ocorre quando o LLM não dispõe do conhecimento adequado para responder ao usuário devido a alguma falha ou carência em seu treinamento, mas gera alguma resposta para não desagradar o interlocutor.

A seguir serão descritos alguns detalhes dos principais grandes modelos de linguagem disponíveis publicamente, inclusive em versões para uso gratuito.

2.2.1 ChatGPT

O ChatGPT (GPT04/GPT-4o, *Chatbot-Generative Pre-Trained Transformer*, ou Assistente Conversacional com Transformador Generativo Pré-Treinado), lançado em 2022, foi o primeiro assistente conversacional dotado de um grande modelo de linguagem que foi disponibilizado ao público, permitindo conversação natural e contextual, com acesso via navegador ou aplicativo próprio. Ele foi desenvolvido pela Open AI e sua arquitetura se baseia num modelo *Transformer* denso, autorregressivo e multimídia, com mecanismo de atenção (*self-attention*) que identifica as partes relevantes do texto para gerar respostas coerentes. Seu treinamento é supervisionado, com realimentação humana e reforço (*Reinforcement Learning from Human Feedback*, RLHF), sendo feito a partir de dados públicos da web, licenciados, multimídia (imagens e áudio) mais código computacional e matemático. Sua janela de contexto, ou seja, o número máximo de *tokens* que o modelo pode processar simultaneamente na entrada e na saída,

varia de aproximadamente 128k até 1M tokens, dependendo da versão. Suas principais características são conversação avançada, dados multimídia (texto, imagem e áudio), geração de código de programação e análise de documentos. Seu principal foco é assistência geral de inteligência artificial e conversacional. Contudo, o ChatGPT é sujeito à ocorrência de “alucinações” (Abbasi et al., 2026; Open AI, 2024).

2.2.2 CoPilot

Já o CoPilot, lançado em 2023 sob o nome de *Bing Chat*, também é baseado em modelos GPT (*Transformer*) e foi desenvolvido pela Open AI e Microsoft para ser integrado aos demais aplicativos desenvolvidos por esta empresa. Ele utiliza uma versão do modelo Prometheus, base GPT-4 da OpenAI, e foi treinado com dados públicos e licenciados, bem como grandes repositórios de código de programação. Sua janela de contexto pode alcançar centenas de milhares de *tokens*, conforme a versão. Suas características principais são a integração com a plataforma Windows, Microsoft 365 e ambientes de desenvolvimento integrado (*Integrated Development Environment*, IDE), além de geração e revisão de código computacional. Seu foco primário está na produtividade empresarial e programação auxiliada. Como os outros grandes modelos de linguagem, também pode sofrer de “alucinações” (Guimarães et al., 2025).

2.2.3 Gemini

A Google lançou o Bard em 2023, o qual foi posteriormente rebatizado como Gemini. Ele usa um modelo *Transformer* multimodal e variantes de mistura de especialistas. Seu treinamento multimodal inclui dados de texto, imagens, áudio e vídeo. Pode ser integrado ao Google Workspace, Android e demais serviços Google. Sua janela de contexto varia entre 1 e 2 milhões de tokens. Suas principais características são a multimodalidade nativa e contextos longos (Gemini Team, 2025).

2.2.4 Claude

O Claude, lançado em 2023, usa arquitetura *Transformer* otimizada para contexto longo e foi desenvolvido pela empresa Anthropic. Os dados usados em seu treinamento são provenientes da *web*, código computacional, conjuntos de dados com curadoria e dados de referência humana. Ele incorpora a Constitutional AI, metodologia de treinamento e alinhamento de

modelos de inteligência artificial criada pela Anthropic para garantir que sistemas de inteligência artificial se comportem de forma segura, útil e alinhada com valores humanos e que reduz necessidade de RLHF. Sua janela de contexto pode chegar até um milhão de *tokens* nas versões mais recentes. É um assistente de inteligência artificial para uso geral, seguro e orientado ao raciocínio, e especializado em textos longos (Anthropic, 2024).

2.2.5 DeepSeek

O modelo DeepSeek foi desenvolvido pela empresa chinesa Hangzhou DeepSeek Artificial Intelligence Basic Technology Research Co., Ltd. e disponibilizado em 2025. Seu lançamento chamou grande atenção, pois seu treinamento requereu custos bem menores do que o verificado nos modelos equivalentes de outras empresas de inteligência artificial, em função do uso de um modelo *open-source* mais eficiente, projetado para uso com menores recursos computacionais. Na verdade, ela aproveitou a infraestrutura de sua empresa proprietária, o fundo High-Flyer, que já utilizava grandes clusters de GPU para algoritmos de *trading*. Sua arquitetura também é baseada em *Transformer* e mistura de especialistas com atenção latente em múltiplas cabeças. Seu treinamento abrange aproximadamente 8 trilhões de *tokens* de dados, entre *web*, ciência, matemática e código computacional. Sua janela de contexto abrange até 128k *tokens*. Seu desempenho é bom em tarefas técnicas, raciocínio estruturado e código computacional, mas sua interface é menos intuitiva para usuários não especializados (DeepSeek-AI, 2025).

2.3 Aplicação dos Grandes Modelos de Linguagem na Siderurgia

O sucesso dos grandes modelos de linguagem naturalmente incentivou os especialistas siderúrgicos a utilizar esse novo recurso nos temas de seu interesse, incluindo o assunto aqui tratado. Nos primórdios dessa nova tecnologia, as respostas dos modelos eram muito simplificadas ou mesmo meras “alucinações”. Contudo, a evolução ocorrida nesses recursos de inteligência artificial ao longo dos últimos anos alterou essa situação, transformando-os em ferramentas bastante úteis. Por outro lado, esses modelos naturalmente não consideram o conhecimento privado disponível pelo seu usuário ou instituição onde ele trabalha (Schneider; Meske; Kuss, 2024). Uma consulta a um LLM que envolva tal conhecimento privado pode igualmente gerar “alucinações”.

Uma solução para esse problema está na execução de um ajuste fino na LLM, ou seja, na atualização do modelo por meio da incorporação de documentos privados, o que melhora seu desempenho nessas áreas específicas, sem requerer o enorme esforço e investimento financeiro necessários para um modelo fundacional original (Klesel; Wittman, 2025).

Uma segunda possibilidade está no uso da chamada Geração Aumentada por Recuperação (*Retrieval Augmented Generation*, RAG), ou seja, a criação de uma base de dados separada com informação vetorizada (*embedding*) (Klesel; Wittman, 2025). Esta alternativa requer ainda menos recursos do que um ajuste fino de LLM (Balaguer 2024). Mas sua implantação ainda não é simples e é necessário dispor de servidores poderosos para que essa abordagem tenha bom desempenho.

Finalmente, a terceira opção está no uso dos grandes modelos de linguagem públicos, mas adotando-se a chamada engenharia de *prompts*, ou seja, a adoção de abordagens mais complexas e estruturadas de diálogo com esses modelos, e não a mera formulação de perguntas diretas. A Tabela 1 mostra as principais características das três abordagens citadas acima.

Tabela 1: Comparação entre as abordagens de *Fine Tuning*, RAG e a Engenharia de *Prompts*.

Característica	Fine Tuning	RAG	Engenharia de Prompts
É possível fazer uso de fontes externas de conhecimento?	x	√	x
Minimiza alucinações?	√	√	√
Requer dados específicos de um domínio?	√	x	x
É adequado para dados dinâmicos?	x	√	x
Ele oferece uma clara interpretabilidade dos resultados?	x	√	x
Baixa utilização de recursos	x	x	√
Rápida implantação	x	x	√

Fonte: Bhavsar (2024)

Um exemplo extremamente complexo sobre o uso de LLM na área siderúrgica foi o desenvolvido por Tian (Tian et al., 2025), a ponto de poder ser considerado um paradigma atual sobre o uso da inteligência artificial no desenvolvimento de novos tipos de aço. Esses pesquisadores desenvolveram uma LLM específica para aços, denominada SteelBERT, a qual

foi pré-treinada com um conjunto estruturado de artigos técnicos e científicos (*corpus*) abrangendo 4,2 milhões de resumos sobre ciências dos materiais e 55.000 textos integrais versando sobre aços. Esse *corpus* inclui informações sobre composição química, tratamento termomecânico, microestrutura, tratamentos térmicos e propriedades mecânicas obtidas. A inclusão dos textos integrais (não apenas resumos) é crucial para capturar as relações complexas entre variáveis metalúrgicas. Os artigos completos foram acessados nos formatos de texto simples, linguagem de marcação extensível (*Extensible Markup Language*, XML) ou de hipertexto (*HyperText Markup Language*, HTML) por meio das interfaces de programação de aplicações (*Application Programming Interface*, API) da Elsevier e da extração automatizada de dados de páginas da internet (*web scraping*) no caso de periódicos publicados pelas editoras Springer, ASME e MDPI.

O SteelBERT se baseia na arquitetura *Transformer* (estilo BERT) e utiliza um mecanismo de atenção desacoplada (*disentangled attention*) que separa a representação do conteúdo semântico e da posição, melhorando a capacidade do modelo em captar relações técnicas longas e dependências complexas típicas das relações metalúrgicas (Tian et al., 2025). Isso viabiliza a previsão das propriedades dos aços a partir das informações de composição química e processo metalúrgico fornecidas, usando-se linguagem natural. Os resultados obtidos mostraram que o SteelBERT apresentou melhor desempenho que modelos generalistas, a especialização do *corpus* técnico melhorou significativamente a capacidade de previsão do modelo e que este captou as relações metalúrgicas implícitas descritas na literatura. Por outro lado, o SteelBERT depende da qualidade do *corpus*, há risco de viés para os tipos de aço mais estudados, como os estruturais e outros produtos *commodity*, e seu desempenho é inferior aos modelos metalúrgicos convencionais quando se requer resultados com alta precisão.

A grande complexidade do SteelBERT é mais um argumento a favor de avaliar o que os atuais modelos de linguagem de grande porte públicos e gratuitos podem oferecer na área de desenvolvimento de aços. Isso vale particularmente para os chamados aços estruturais, um produto *commodity* que vem sendo intensamente estudado há várias décadas. Há grande chance de que o conhecimento disponível para a produção desse material específico já tenha sido incorporado aos LLMs de acesso público, dispensando a necessidade do desenvolvimento de modelos especializados neste tipo de material. Por outro lado, é interessante desenvolver uma metodologia confiável para avaliação do desempenho desses LLMs quando aplicados na resposta a consultas técnicas sobre aços estruturais microligados, para se determinar qual deles oferece respostas de alta qualidade, completas e sem a ocorrência de alucinações.

2.4 Metodologias para Avaliação do Desempenho de LLMs

A competição entre várias organizações no desenvolvimento de várias arquiteturas de *chatbots* usando LLMs, com acesso público e gratuito, criou a necessidade de se desenvolver metodologias para se proceder à avaliação sistemática e neutra de seu desempenho quando aplicadas a objetivos específicos. As diferenças fundamentais na filosofia de seus projetos criam dúvidas sobre seus pontos fortes e fracos em aplicações práticas (Abbasi et al., 2026).

As abordagens atuais para avaliação de LLMs apresentam diversas limitações que impedem uma comparação significativa. Conjuntos de dados inconsistentes constituem um desafio para uma comparação direta, uma vez que o desempenho dos LLMs pode variar de forma significativa conforme as características dos dados. Diferentes métricas de avaliação complicam essa comparação, dependendo da ênfase dada à precisão, fluência, coerência ou medidas específicas para o caso (Abbasi et al., 2026). Finalmente, é necessário considerar adequadamente as técnicas de engenharia de *prompts*, para que se leve em conta a significativa influência dessa formatação do desempenho do modelo (Liu et al., 2023).

No caso particular da avaliação de desempenho dos LLMs, pode ser muito útil usar o *role prompt*, ou seja, informar ao modelo qual o papel que ele deverá assumir. Por exemplo, um LLM pode assumir o papel de um examinando ao responder perguntas ou consultas sobre um dado tema, e outro LLM assume o papel de examinador ao analisar essas respostas. A definição de uma *persona* ajuda o LLM a se ajustar à tarefa a ser executada, definindo seu estilo, contexto de raciocínio, foco técnico e nível de detalhe. No caso do LLM examinando deve-se incluir o público alvo e, para o LLM examinador, os critérios de avaliação da resposta. A atribuição de um papel não modifica o conhecimento do modelo, mas sim orienta a seleção de informações, o estilo da resposta e a estrutura da explicação (Chen 2025; Sahoo 2025).

A avaliação humana tornou-se o padrão na avaliação do desempenho dos LLMs e, de fato, vem sendo usada cada vez mais em seu treinamento. Contudo, alguns autores (Hosking, Blunsom e Bartolo, 2024) acreditam que as notas concedidas usando essa abordagem podem ser subjetivas e levar a vieses indesejáveis. Os avaliadores humanos não são infalíveis e apresentam tendência a procurar atalhos para tornar sua tarefa mais fácil e menos tediosa, fazendo com que seu julgamento se concentre em características superficiais, tais como fluência e complexidade linguística, desprezando aspectos que requerem maior esforço para serem analisados – como, por exemplo, fatos reais e que podem ser verificados (Ipeirotis, Provost e Wang, 2010). Além disso, foi verificado que respostas mais assertivas tendem a ser preferidas pelos avaliadores humanos, fazendo com que outras características, como inconsistências e factualidade,

sejam sub-representadas. Uma vez que os modelos estão se tornando cada vez mais poderosos, a diferença entre qualidade percebida e a utilidade real da resposta se tornarão cada vez mais importantes. As soluções propostas para se mitigar esse problema, ainda que parcialmente, é a formação de um conjunto selecionado de avaliadores treinados e motivados, ou efetuar avaliações múltiplas cuidadosamente agregadas (Hosking, Blunsom e Bartolo, 2024). Por outro lado, é necessário levar em conta que a garantia da qualidade das avaliações humanas requer algum tipo de remuneração ou recompensa para os avaliadores, o que dificulta a adoção desta abordagem em trabalhos acadêmicos, onde não há recursos para tal. Finalmente, é necessário considerar que o trabalho associado às avaliações humanas pode ser demorado e dispendioso em função da complexa logística envolvida (Wang et al, 2023; Bhavsar, 2025).

Esta situação acabou por criar condições para se utilizar os próprios grandes modelos de linguagem como avaliadores do desempenho de outros modelos similares em situações específicas, dentro da abordagem *LLM-as-a-Judge* (LLM como Juiz), particularmente quando as métricas para se avaliar as respostas dos modelos não são numéricas. Outras vantagens são a rapidez e baixos custos associados a esta abordagem, além da minimização dos vieses típicos das avaliações humanas (Wang et al, 2023; Bhavsar, 2025).

Por outro lado, o uso de LLMs como avaliadores também tem suas desvantagens. Um mesmo LLM não pode avaliar suas próprias respostas. Isso é óbvio, pois neste caso ocorre um efeito de “nepotismo” que faz com que ele conceda melhores avaliações a elas, mesmo que as respostas de outros modelos sejam tão boas ou mesmo melhores do que elas. Isso ocorre devido à familiaridade do LLM com seu próprio estilo de escrita. Respostas com melhor estética, ainda que incompletas, tendem a ser mais bem avaliadas do que respostas completas, mas redigidas de forma simples. Outro problema é a confusão que os LLMs fazem entre qualidade de informação e o tamanho das respostas (Wang et al, 2023; Bhavsar, 2025).

Há ainda o viés de posição: grandes modelos de linguagem tendem a favorecer a primeira resposta para uma mesma questão em detrimento das seguintes, mesmo que esta apresente melhor qualidade que a primeira. Uma situação semelhante ocorre durante a avaliação de longas respostas: os LLMs tendem a exibir um padrão de atenção em forma de “U”, concedendo peso desproporcional à informação no início e no fim do *prompt*, enquanto eles potencialmente desconsideram detalhes cruciais na metade. Ou seja, do ponto de vista da psicologia cognitiva:

- O conteúdo no início do *prompt* recebe alta atenção (40%, “viés de primazia”)
- O conteúdo no final do *prompt* recebe foco mais intenso (40%, “viés de recência”)

- Informações importantes nas seções intermediárias do *prompt* são subestimadas ou negligenciadas (20% de atenção);
- Esse viés afeta a qualidade da avaliação, independentemente da importância real de cada seção.

Esse problema pode ser minimizado através da alteração da sequência das respostas apresentadas ao LLM examinador (Wang et al, 2023; Shi et al, 2025; Bhavsar, 2025).

3 METODOLOGIA

3.1 Premissas Básicas

Foi elaborado um questionário sobre os principais conceitos metalúrgicos e práticas industriais sobre o processamento por tratamento termomecânico (mais conhecido pelo jargão “laminação controlada”) de aços planos microligados estruturais. Ele era constituído por um preâmbulo para servir de motivação para o LLM candidato e dez questões sobre o tema em questão. Seu texto integral se encontra no Apêndice A. Optou-se por redigi-lo em inglês, língua predominante da literatura e informações disponíveis sobre o assunto. Além disso, espera-se que o uso desse idioma evite uma possível perda de desempenho dos LLMs decorrente da tradução de sua língua nativa para o português, e vice-versa. Ao responder esse tipo de questão em português, muitas vezes os LLMs misturam o jargão em inglês no texto.

Foram escolhidos neste trabalho assistentes conversacionais (*chatbots*) de inteligência artificial mais famosos, de acesso público e gratuito, baseados em grandes modelos de linguagem, tanto para responder o questionário sobre a laminação controlada, como para avaliar as respectivas respostas, dentro da abordagem *LLM-as-a-Judge*: Chat GPT (<https://chatgpt.com>), Copilot (<https://copilot.microsoft.com>), Gemini (<https://gemini.google.com>), Claude (<https://claude.ai>) e Deep Seek (<https://www.deepseek.com>). Conforme descrito anteriormente, todos esses sistemas se baseiam na arquitetura *Transformer*, que utiliza mecanismos de atenção para modelar relações entre *tokens* e as sequências de texto.

3.2 Procedimento

Foram anotadas as datas e horários em que as perguntas do questionário foram fornecidas e respondidas, bem como o momento em que essas respostas foram corrigidas. O ciclo de obtenção e avaliação das respostas ao questionário foi o mais rápido possível, de forma a manter constante, tanto quanto possível, o desempenho dos diversos LLMs, minimizando o risco de eles serem expostos a novos dados ou atualizados nesse meio tempo.

Antes que o LLM respondesse ao questionário, foi-lhe perguntado a ele qual era a sua versão atual e data de sua última atualização. Infelizmente, no modo *chatbot*, eles não informam o valor do parâmetro temperatura, o que impediu uma análise mais detalhada sobre a reprodutibilidade das respostas ao questionário e as respectivas correções.

A seguir, foi introduzido o preâmbulo, onde foi atribuído ao LLM em questão o papel de examinando e foi-lhe atribuída a tarefa de responder a um questionário técnico que o avaliaria para um cargo de supervisor técnico no setor de laminação a quente de uma usina. Também foi informado a ele a maneira com que ele deveria responder às questões. A seguir, as dez questões lhe foram apresentadas, uma a uma, para que o LLM elaborasse a correspondente resposta. Esta era então copiada para um arquivo em Word. Seguia-se então a apresentação da próxima pergunta, até o final do questionário.

A sequência em que os *chatbots* responderam todo o questionário foi definida de forma algo arbitrária, baseada na ordem cronológica de seu lançamento comercial, sendo definida como ChatGPT – CoPilot – Gemini – Claude – DeepSeek.

A fase de avaliação das respostas elaboradas pelos LLMs iniciou-se assim que todos eles haviam respondido ao questionário. Cada LLM avaliou as respostas geradas pelos demais e, obviamente, ele não avaliou suas próprias respostas, para evitar a ocorrência de “nepotismo” e favoritismo. O LLM que gerou a resposta tende a reconhecer seu estilo e concederá melhor avaliação a ela (Bhavsar 2025). Também aqui, antes que o LLM avaliasse as respostas elaboradas pelos outros LLMs, era perguntado a ele qual era a sua versão atual e a data de sua última atualização.

No início da avaliação de cada questionário pelas LLMs, era introduzido um preâmbulo, que solicitava ao *chatbot* que o assumisse o papel de um examinador, com mestrado e doutorado em metalurgia, e assumisse a missão de avaliar tecnicamente as respostas fornecidas ao questionário. A seguir eram informados os parâmetros de avaliação a serem considerados para cada resposta do questionário, devendo ser atribuída nota para cada um deles. Esse texto encontra-se no Apêndice B. A fim de evitar alteração na severidade de avaliação que ocorre quando um LLM analisa respostas diferentes para uma mesma questão, todas respostas elaboradas por uma dada LLM, naturalmente com os correspondentes enunciados, foram submetidas sequencialmente ao LLM avaliador para tentar evitar o chamado viés de posicionamento (Bhavsar 2025). Após cada apresentação do par pergunta/resposta e a correspondente correção, era feita uma pausa e o texto era copiado para um arquivo em Word. Finalmente, ao final da avaliação das questões, era perguntado ao LLM avaliador se o candidato reunia condições técnicas para assumir o cargo pretendido.

Os critérios que foram adotados para a avaliação das questões estão listados a seguir:

- Correção técnica;
- Relevância;

- Coerência metalúrgica;
- Coerência de raciocínio;
- Coerência lógica;
- Uso correto da terminologia técnica;
- Clareza, profundidade, organização;
- Objetividade, concisão;
- Completude;
- Ausência de “alucinações”.

Cada LLM candidato foi avaliado de acordo com a média das notas das provas concedidas pelos demais LLMs avaliadores. Todos os itens de avaliação tinham o mesmo peso. A partir dessa nota média foi estabelecido um ranking dos LLMs avaliados.

A ordem dos *chatbots* durante a fase de correção de suas respostas ao questionário foi a mesma adotada durante a fase de elaboração de suas respostas: ChatGPT – CoPilot – Gemini – Claude – DeepSeek. Isso fez com que o desempenho dos primeiros candidatos na sequência fosse eventualmente comentado pelo LLM avaliador durante a correção dos questionários dos LLMs candidatos que se encontravam em lugar posterior na sequência de correções, mesmo em sessões diferentes, realizadas com alguns dias de diferença.

3.3 Disponibilização dos Dados Usados Neste Trabalho

Os textos integrais das questões, respostas e correções elaboradas pelos *chatbots* não puderam ser incluídos aqui, dado o seu grande tamanho, que alcançou um total de aproximadamente 1.500 páginas. Mas os eventuais interessados podem encontrar a íntegra não-editada desses textos no GitHub, na forma de arquivos PDF, acessando o endereço <https://github.com/antonigorni/TCC---ICMC>.

Vale lembrar que, por se tratar de textos gerados por inteligência artificial, eles podem conter informações e conceitos metalúrgicos incorretos e não devem ser usados, em hipótese nenhuma, como única fonte de conhecimento sobre esses tópicos.

4 ANÁLISE E RESULTADOS

4.1 Respostas dos LLMs aos Questionários

As provas foram aplicadas aos cinco LLM's no dia 9 de abril de 2026. No momento da geração das respostas ao questionário eles operavam sob as seguintes versões:

- Chat GPT: *model GPT-5, current version GPT-5.2, last training update, August 2025;*
- Gemini: *model Gemini 3 Flash, atualizado de forma contínua;*
- Claude: *model Sonnet 4.5, last update at the end of January 2025;*
- DeepSeek: *version DeepSeek-V3, variant DeepSeek-V3-0324, last update at March 24, 2025.*

Já o CoPilot informou que não dispunha de um número de modelo ou versão, como tradicionalmente há para as demais plataformas, mas que sua atualização ocorria de forma contínua.

As respostas foram geradas rápida e ininterruptamente para todos os LLMs, exceto o Claude que, em alguns casos, suspendeu o serviço gratuito após responder a quarta questão e só retomou as atividades quatro horas depois. Geralmente não foram incluídos gráficos nas respostas; em alguns poucos casos, foram gerados gráficos esquemáticos usando caracteres.

A Tabela 2 mostra o número de caracteres de cada resposta para as questões da prova que foi redigida pelos LLMs candidatos. Pode-se observar que o Gemini gerou respostas cujo tamanho e desvio padrão foram mínimos, ao contrário do Claude. Também o ChatGPT e CoPilot elaboraram respostas com tamanho abaixo da média, enquanto ocorreu o contrário com o DeepSeek. Esse aumento de tamanho corresponde ao maior detalhamento verificado nas respostas geradas pelo Claude e DeepSeek.

Por sua vez, a Tabela 3 mostra, para cada LLM candidato, o número total de caracteres, tempo dispendido na geração, velocidade e número total de páginas ocupadas pelas respostas às questões. O tempo de execução é aproximado, já que ele incluiu a transcrição manual das respostas elaboradas pelos LLMs para um arquivo Word. Os candidatos mais prolixos, Claude e DeepSeek, também foram os mais rápidos; curiosamente, o Gemini, o mais lacônico, apresentou a menor velocidade na geração de respostas. Note-se que o número de páginas das respostas foi surpreendentemente alto; o LLM mais lacônico, por assim dizer, foi o Gemini, com 23 páginas, e o Claude o mais prolixo, com 58. Esse fato evidencia claramente que as respostas

foram geradas por LLMs, pois os números de páginas aqui obtidos são irrealistas para candidatos humanos, limitados pelo tempo para execução da prova e pelo seu conhecimento disponível. As cinco respostas ao questionário alcançaram um total de 204 páginas o que, na prática, torna inviável que sua correção seja feita por humanos, a menos que disponham do tempo necessário e interesse em efetuar essa atividade, algo muito improvável se feito *pro bono*.

Tabela 2: Número de caracteres (exceto espaços) das respostas às questões geradas pelos LLMs candidatos. DPR significa Desvio Padrão Relativo em relação à média, expresso em percentagem.

Questões → Candidatos ↓	1	2	3	4	5	6	7	8	9	10	Média	DPR [%]
ChatGPT	3488	3577	2912	3085	3845	2043	3236	2994	4362	4086	3363	20%
CoPilot	3333	4771	4729	5198	5777	3274	4841	3001	6634	6189	4775	26%
Gemini	2469	3132	3104	3310	2831	2010	2628	2405	2972	2925	2779	14%
Claude	3770	5534	7645	12319	5740	5567	8358	8030	13630	14498	8509	44%
Deep Seek	2912	4134	8269	9636	6454	5224	6138	5312	9414	8075	6557	34%
Média	3194	4230	5332	6710	4929	3624	5040	4348	7402	7155	5196	28%
DPR [%]	16%	23%	47%	61%	31%	47%	46%	54%	57%	64%		

Tabela 3: Número total de caracteres (exceto espaços), tempo dispendido, velocidade e número de páginas ocupadas pelas respostas às questões geradas pelos LLMs candidatos.

Candidato	Caracteres	Tempo Dispendido [min]	Velocidade [caracteres/min]	Páginas
ChatGPT	33628	18	242	30
CoPilot	47747	31	214	41
Gemini	27786	18	165	23
Claude	85091	20	682	58
Deep Seek	65568	27	349	52
Média	51964	23	330	Total
DPR [%]	45%	26%	63%	204

4.2 Correções das Respostas ao Questionário Geradas pelos LLMs Candidatos

Os conjuntos das respostas elaboradas pelos cinco LLMs foram corrigidos pelos quatro LLMs avaliadores restantes, totalizando vinte conjuntos de correções, entre os dias 11 e 15 de abril de 2026. Os dados sobre modelos e versões dos LLMs foram idênticos aos já relatados na seção anterior. As correções foram feitas de forma rápida e ininterrupta, exceto, novamente, no caso do Claude, que sistematicamente interrompia a correção antes de chegar ao fim do questionário, só permitindo a retomada da tarefa cerca de quatro horas depois. Isso causou algum

transtorno pois, a cada retomada de atividade, era necessário informar novamente o Claude sobre o *prompt* inicial de avaliação, bem como as avaliações das questões que já haviam sido feitas. Por sua vez, numa de suas correções, o ChatGPT e CoPilot subitamente abandonaram o procedimento que havia sido previamente informado após já terem corrigido algumas respostas, passando a adotar seus próprios critérios. Curiosamente, nesse momento o ChatGPT espontaneamente passou a se expressar em português. Era então necessário interromper a atividade e reapresentar o *prompt* inicial para que a metodologia original de correção fosse retomada.

A Tabela 4 mostra que o *ranking* dos LLMs em termos do tamanho das correções, expressos em número de caracteres, foi um pouco diferente ao que já havia sido observado no caso das respostas às perguntas. Claude manteve a liderança, sendo seguido de longe por ChatGPT e DeepSeek. De forma análoga, Claude se caracterizou por apresentar as correções mais detalhadas, ao contrário do Gemini, que manteve seu perfil lacônico.

Tabela 4: Número total de caracteres (exceto espaços), das correções às respostas ao questionário geradas pelos LLMs avaliadores.

Questões → Avaliador ↓	1	2	3	4	5	6	7	8	9	10	Média	DPR [%]
ChatGPT	14923	16861	17507	17316	18034	18449	20458	21764	22154	17180	18465	12%
CoPilot	11995	13556	13842	16091	15050	14797	16355	15293	15716	15679	14837	9%
Gemini	9273	9963	10155	10230	10508	10216	10751	10778	10665	11559	10410	6%
Claude	21711	29424	35085	38347	30757	31604	28678	31453	32039	32885	31198	14%
DeepSeek	12199	16779	17934	18901	18707	19031	17406	17277	17477	17220	17293	11%
Média	14020	17317	18905	20177	18611	18819	18730	19313	19610	18905		
DPR [%]	34%	42%	51%	53%	40%	42%	35%	41%	41%	43%		

Por sua vez, a Tabela 5 mostra que o número de páginas das correções efetuadas, que também incluiu o enunciado das perguntas e as respectivas respostas, foi de algumas centenas para cada LLM, atingindo um total de quase 1.500 páginas. Isso é mais uma evidência da prioridade dos LLMs em maximizar a completude de suas correções, frequentemente indo muito além do que havia sido solicitado nos enunciados das questões. Não há uma correspondência perfeita entre os números totais de caracteres e de páginas em função dos diferentes estilos de diagramação e orientação de página nos textos gerados pelos diferentes LLMs. Além disso, o ChatGPT fez uma série de comentários adicionais sobre o candidato ao final de cada correção das respostas do questionário, cujos números de caracteres não foram computados dentro das correções das questões propriamente ditas. O Claude apresentou velocidade máxima de

resposta, apesar do maior tamanho de suas correções, seguindo-se o DeepSeek, ChatGPT, Gemini e CoPilot.

Tabela 5: Número total de caracteres (exceto espaços), tempo dispendido, velocidade e número de páginas ocupadas pelas correções às respostas das perguntas geradas pelos LLMs avaliadores.

Avaliador	Caracteres	Tempo Dispendido [min]	Velocidade [caracteres/min]	Páginas
ChatGPT	184646	116	1592	340
CoPilot	148374	112	1325	282
Gemini	104098	72	1446	249
Claude	311983	104	3000	338
DeepSeek	172931	88	1965	275
Média	184406	98	1865	Total
DPR [%]	42%	19%	36%	1484

CoPilot e DeepSeek, incluíram *links* da internet para artigos técnicos em algumas de suas respostas e correções, atuando como se fossem um motor de busca; provavelmente o conteúdo deles foi consultado para gerar suas saídas. Por sua vez, Claude só citou os nomes dos autores envolvidos nos assuntos tratados ou de seus respectivos livros, sem *links* externos. Em suas avaliações, Claude penalizou a nota dos LLMs candidatos quando verificou o uso de *links* de internet nas respostas, argumentando que eles não deveriam ser incluídos em avaliações técnicas formais. Mas, no caso específico da correção das respostas do DeepSeek, Claude ainda alegou que as informações muito específicas presentes nesses artigos não estabeleceram uma resposta geral, conforme solicitado no enunciado da pergunta.

Em algumas de suas respostas e correções a respostas do questionário elaboradas por outros LLMs, CoPilot e DeepSeek geraram uma espécie de “prólogo”, onde era esquematizada sua linha de raciocínio antes de efetivamente responder à questão ou corrigir a resposta à questão. Curiosamente, o DeepSeek, ao corrigir das respostas do questionário elaboradas pelo CoPilot, criticou esse procedimento, alegando que isso não era um correto por parte do candidato.

4.3 Avaliação das Respostas ao Questionário

O procedimento de avaliação adotado pelos diferentes LLMs foi bastante semelhante entre eles: a cada resposta corrigida do questionário era gerada uma tabela contendo as notas correspondentes aos itens solicitados de avaliação solicitados no *prompt*, justificativas para as

notas dadas, um resumo sobre a avaliação da questão e uma lista dos pontos fortes e fracos da resposta, incluindo uma análise para verificar se a qualidade da resposta seria compatível com um candidato a um posto de supervisão industrial. Alguns LLMs candidatos adotaram um estilo muito informal nas respostas, o que gerou críticas por alguns LLMs avaliadores, que questionaram a seriedade do candidato.

A Tabela 6 apresenta as notas concedidas pelos LLMs avaliadores aos LLMs candidatos, conforme os itens que haviam sido solicitados no *prompt* inicial, bem como sua avaliação global. Constata-se que a amplitude entre as notas das questões foi relativamente alta, tendo oscilado entre 3,6 e 5,0 e apresentado desvio padrão relativo entre 4% e 13%.

Tabela 6: Notas concedidas pelos LLMs avaliadores para as respostas ao questionário que foram fornecidas pelos LLMs candidatos, de acordo com os itens de avaliação considerados, e sua média global.

Candidatos → Critérios ↓	ChatGPT	CoPilot	Gemini	Claude	DeepSeek	Média	DPR [%]
Correção Técnica	4,3	4,1	4,1	4,8	4,1	4,3	7%
Relevância	4,6	4,7	4,8	5,0	4,9	4,8	4%
Coerência Metalúrgica	4,5	4,2	4,3	4,8	4,2	4,4	6%
Coerência de Raciocínio	4,5	4,4	4,8	4,9	4,5	4,6	5%
Coerência Lógica	4,6	4,5	4,7	4,8	4,3	4,6	4%
Terminologia Técnica	4,6	4,5	4,4	5,0	4,6	4,6	5%
Clareza	4,1	4,2	4,9	5,0	4,8	4,6	9%
Objetividade	4,1	3,8	4,2	4,2	3,7	4,0	6%
Compleitude	3,6	4,0	3,9	5,0	4,5	4,2	13%
Ausência de Alucinações	4,9	4,3	4,5	4,7	4,2	4,5	6%
Média	4,36	4,26	4,45	4,82	4,38		
DPR [%]	8%	6%	8%	5%	8%		

O item que foi pior avaliado, para todos os LLMs candidatos, foi Objetividade, com média geral de 4,0, em especial no caso de DeepSeek e CoPilot. Provavelmente isso se deve ao detalhamento exagerado em suas respostas. Paradoxalmente, a segunda pior nota geral (4,2) foi no item Compleitude, onde vários LLMs candidatos foram criticados por seus avaliadores por não incluírem em suas respostas informações que estes achavam importantes, alegando que isso fazia com que suas respostas parecessem que haviam sido extraídas de livros-texto, pois não mencionavam detalhes práticos ou industriais. Outra crítica que apareceu com frequência foi a falta de abordagem quantitativa nas respostas às questões ou a apresentação de equações que

descrevem os fenômenos metalúrgicos. A análise do item Completude parece ter tido caráter bem subjetivo, pois suas notas variaram desde 3,6 para o ChatGPT até 5,0 para o Claude, apresentando ainda desvio padrão máximo, de 13%. De forma geral, a avaliação da Completude apresentou correlação com o tamanho das respostas. O item Correção Técnica apresentou a terceira pior nota (4,3), com CoPilot, Gemini e DeepSeek apresentando o pior desempenho (4,1). Seguiu-se a Coerência Metalúrgica (4,4), onde apenas o Claude teve boa avaliação (4,8). Já o item Relevância teve a melhor avaliação, com média geral igual a 4,8 e desvio padrão mínimo, de 4%. Os demais itens da avaliação apresentaram notas intermediárias.

A variação nos valores das médias finais dos LLMs candidatos não foi alta, oscilando entre 4,4 e 4,8 para um máximo de 5,0, com desvio padrão relativo variando entre 5% e 8%. Como já seria de se esperar, em função do alto volume de informações e um mínimo de alucinações em suas respostas, Claude foi o LLM que se saiu melhor, recebendo nota 4,82. Curiosamente, Gemini, o LLM menos prolixo da série aqui avaliada, ficou em segundo lugar, com nota 4,45. O terceiro lugar ficou com DeepSeek, com nota 4,38, praticamente empatado com ChatGPT, que recebeu nota 4,36. O último lugar coube a CoPilot, com nota 4,26.

As notas concedidas pelos LLMs avaliadores para os LLMs candidatos estão listadas na Tabela 7. Pode-se verificar um grau significativo de subjetividade entre os LLMs avaliadores, que decorre da média global atribuída por eles aos vários LLMs candidatos – de 3,7 (Claude) para 4,9 (Gemini).

Tabela 7: Notas globais concedidas pelos LLMs avaliadores aos LLMs candidatos.

Candidatos → Avaliadores ↓	ChatGPT	CoPilot	Gemini	Claude	DeepSeek	Média	DPR [%]
Chat GPT	-	4,3	4,4	4,7	4,3	4,4	4%
CoPilot	4,7	-	4,6	4,7	4,6	4,7	1%
Gemini	4,9	4,9	-	4,9	4,9	4,9	0%
Claude	3,4	3,5	4,0	-	3,7	3,7	7%
Deep Seek	4,4	4,3	4,8	5,0	-	4,6	7%
Média	4,4	4,3	4,5	4,8	4,4		
DPR [%]	15%	14%	8%	3%	12%		

Claude foi o avaliador mais severo, o que não é inesperado, uma vez que ele gerou as respostas mais abrangentes ao questionário em comparação às geradas pelos demais LLMs candidatos, com notas variando entre 3,4 e 4,0. Por outro lado, Gemini constituiu uma exceção, tendo recebido a maior nota concedida pelo Claude, ainda que não muito alta (4,0), mesmo

tendo apresentado respostas com tamanho mínimo dentre todos os LLMs candidatos. O ChatGPT foi o segundo LLM avaliador mais severo. Os demais foram mais complacentes, principalmente o Gemini, que atribuiu nota 4,9 para todos os LLMs candidatos, zerando na prática o valor de seu desvio padrão relativo, o que levanta dúvidas sobre seus critérios de correção. Situação similar ocorreu com o CoPilot, onde o valor desse parâmetro também foi extremamente baixo (1%). Já as notas concedidas pelo Claude e DeepSeek apresentaram máximo valor de desvio padrão relativo (7%), o que parece indicar correções mais criteriosas. Curiosamente, DeepSeek atribuiu ao Claude nota máxima (5,0) para todos os quesitos de avaliação o que, naturalmente, impactou sua nota final.

Finalmente, as recomendações finais para a contratação (ou não) dos LLMs candidatos pelos LLMs avaliadores está na Tabela 8, onde se vê que a contratação dos candidatos Claude e Gemini foi unânime entre eles. O ChatGPT ficou em terceira posição, com o Claude restringindo sua contratação que, segundo ele, deveria ocorrer em caráter provisório dentro de um estágio de doze meses na função, com nova avaliação após esse período; já o DeepSeek vetou incondicionalmente sua contratação. Tais recomendações são compatíveis com as notas obtidas pelos LLMs candidatos nas suas respectivas respostas ao questionário. Os demais LLMs candidatos – CoPilot e DeepSeek – foram pior avaliados.

Tabela 8: Veredictos recebidos pelos LLMs candidatos sobre sua contratação como supervisor técnico de uma instalação de laminação a quente.

Avaliadores → Candidatos ↓	ChatGPT	CoPilot	Gemini	Claude	DeepSeek	Veredicto
Chat GPT	-	Sim	Sim	Estagiário	Não	Não
CoPilot	Como Trainee	-	Sim	Exame Oral	Não	Não
Gemini	Sim	Sim	-	Sim	Sim	Sim
Claude	Sim	Sim	Sim	-	Sim	Sim
Deep Seek	Como Trainee	Não	Sim	Exame Oral	-	Não

Curiosamente, mesmo o Claude, tão bem avaliado, apresentou incorreções em suas respostas e correções. Ainda que em apenas dois casos, ele e DeepSeek cometeram erros elementares ao calcular o somatório das notas das respostas a questões.

Da mesma forma como ocorreu em algumas de suas respostas ao questionário, CoPilot e DeepSeek geraram uma espécie de “prólogo” em algumas de suas correções, onde era esquematizada sua linha de raciocínio antes de efetivamente responder à questão.

4.4 Qualidade das Respostas e Avaliações do Ponto de Vista Metalúrgico

De maneira geral, o conhecimento metalúrgico revelado pelos LLMs foi muito volumoso, satisfatório e até surpreendente em relação ao verificado há poucos anos atrás. Contudo, foi verificada a ocorrência de algumas incorreções significativas, tanto nas respostas ao questionário, como em suas correções. Em alguns casos é possível identificar que a origem do problema é humana e deve ter sido absorvida pelos LLMs durante seu treinamento. Em outros observa-se falha efetiva do *chatbot*. Os casos mais significativos estão descritos a seguir:

- Na resposta à questão 2 (distribuição de deformação entre as etapas de esboçamento e acabamento), ChatGPT fez confusão entre três parâmetros importantes que definem os parâmetros de processo da laminação controlada: *Recrystallization Limit Temperature* (RLT), *Recrystallization Stop Temperature* (RST) e *Temperature of No-Recrystallization* (T_{nr}) (Lopez et al, 2022). CoPilot e Claude detectaram esse problema na correção dessa resposta. Acima de RLT ocorre 100% de recristalização da austenita entre os passes, situação em que deve ocorrer a etapa de esboçamento da laminação controlada. Mas, exatamente sob essa temperatura, a recristalização começa a se tornar incompleta; só 95% da microestrutura se recristaliza, ficando os restantes 5% encruados em função da deformação aplicada no passe anterior. Abaixo de RLT a recristalização se torna progressivamente incompleta, até que só ocorra 5% de recristalização da austenita sob RST. Nessa faixa de temperaturas ocorre a fase de espera, ou seja, não se deve laminar a chapa, pois a recristalização incompleta do material torna sua microestrutura heterogênea, o que prejudica a tenacidade (ou seja, resistência ao impacto) do produto final. Abaixo de RST a recristalização da austenita fica totalmente suprimida, podendo-se então continuar a laminação, já que a microestrutura do material se mantém homogênea, apesar de apresentar encruamento progressivo à medida que a laminação prossegue. Por sua vez, em função das condições específicas do ensaio de torção a quente usado na determinação de T_{nr} , não se sabe exatamente a que fração de recristalização ela corresponde, ao contrário de RLT e RST. O conceito de T_{nr} foi o primeiro a surgir e se consagrou significativamente muito antes da definição de RLT e RST, fazendo com que vários autores acabem confundindo esses parâmetros afirmando, por exemplo, que RST e T_{nr} são idênticos, o que não é correto (Stalheim; Glodowski, 2010).

- Ainda na questão 2, a resposta de ChatGPT não considerou o impacto que a espessura do esboço (a chapa que está sendo laminada) na etapa de espera da laminação controlada exerce sobre a produtividade do Laminador de Chapas Grossas, uma falha imperdoável para um candidato que está sendo submetido a uma sabatina para verificar seu preparo técnico visando assumir um posto de supervisão numa unidade industrial de laminação (Stalheim; Glodowski, 2010). Todos os outros LLMs avaliadores (CoPilot, Gemini, Claude e DeepSeek) apontaram essa falha. Quanto maior a espessura do esboço na fase de espera, maior a quantidade de deformação a ser aplicada nele sob baixas temperaturas, situação em que o material está mais duro e torna necessária a aplicação de maior número de passes de laminação sob menores graus de deformação, de forma a não ultrapassar a capacidade mecânica do equipamento. Isso prolonga significativamente a duração da laminação.
- Continuando na questão 2, Gemini notou a ausência de menção ao fator de forma (*shape factor*) na resposta a essa questão elaborada pelo DeepSeek. Esse parâmetro, definido pela razão l_d/h_m entre o comprimento do arco de contato entre esboço e cilindros de laminação, e a média das espessuras de entrada e saída do passe de laminação, quantifica o grau de penetração da deformação até o centro da espessura do laminado. Quanto maior ele for, maior será a probabilidade da ocorrência de caldeamento de vazios e homogeneização da segregação interna na chapa, o que contribui significativamente para sua qualidade (Okumura et al., 1984; Streisselberger; Schwinn; Hubo, 2001). Mas, curiosamente, Gemini não observou essa ausência nas respostas dos outros LLMs, como ChatGPT, CoPilot e Claude.
- Na resposta à questão 3 (condições recomendadas para a fase de esboçamento da laminação controlada), CoPilot afirma que um esboçamento ideal promove a solubilização do Nb. ChatGPT aceitou explicitamente essa afirmação e Gemini não a questionou; por sua vez, Claude e DeepSeek refutaram essa afirmação, alegando que a solubilização do Nb é função primordial do reaquecimento da placa de aço no forno – o que, de fato, é um entendimento consagrado na prática industrial. Mas o fato é que o CoPilot pensou “fora da caixa” e pode ter alguma razão, embora possa ter dado muita importância a essa possibilidade teórica. Ao menos enquanto a temperatura do esboço sob laminação estiver acima da RLT, ainda há condições termodinâmicas para a solubilização do nióbio

continuar a ocorrer durante a fase de esboçamento, pois a cinética de dissolução é lenta demais para promover a solubilização completa dos precipitados de nióbio durante o reaquecimento da placa sob as temperaturas normalmente usadas sob condições industriais, que dificilmente ultrapassam 1250°C (Bernetic et al, 2010). Contudo, como a fase de esboçamento é relativamente curta, durando apenas alguns minutos, pode ser que essa solubilização suplementar seja desprezível em termos práticos.

- Ainda na questão 3, Gemini recomendou que cada passe da laminação de esboçamento apresentasse valor do fator de forma superior a 0,6, para garantir a aplicação de um valor mínimo de deformação que alcance o centro da espessura do laminado, minimizando os defeitos internos decorrentes da solidificação da placa. DeepSeek concordou com essa afirmação quando corrigiu a resposta de Gemini. O princípio é correto, ainda que o valor efetivamente recomendado varie entre 0,8 e 1,0, conforme a fonte (Okumura et al., 1984; Streisselberger; Schwinn; Hubo, 2001). Curiosamente, ChatGPT e Claude “alucinaram” na correção dessa resposta, afirmando que o fator de forma é realmente citado na teoria da laminação, mas que não era um parâmetro primário para definir o esquema de passes na laminação esboçadora, quando ocorre justamente o contrário na moderna prática industrial. Curiosamente, Claude manteve essa opinião ao corrigir as respostas elaboradas pelo Gemini para as primeiras questões, mas mudou de opinião a seguir, afirmando exatamente o contrário a partir da correção da resposta à questão 5. Mas há uma provável causa para isso: durante a correção das respostas do Gemini, o Claude interrompeu suas atividades após a questão 4, só permitindo o retorno dessa atividade algumas horas depois. Ele mudou de opinião sobre o fator de forma após a retomada da correção. Nesse ponto, Claude ainda afirmou que o fator de forma é um parâmetro sofisticado, elogiando sua citação pelo Gemini, e criticou o fato de que tanto o ChatGPT como o CoPilot não o tinham mencionado em suas respostas. Note-se aqui que, nessa oportunidade, o Claude “lembrou” que já tinha corrigido as respostas do questionário desses dois últimos *chatbots*.
- Na resposta à questão 4 (condições recomendadas para a fase de acabamento da laminação controlada), e também em respostas a outras questões, CoPilot recomenda que a temperatura final de laminação fique em torno da temperatura Ar_3 ou um pouco abaixo dela. Essa é a temperatura abaixo da qual a austenita começa a se transformar em ferrita. De fato, a prática industrial consagrada efetivamente consiste em terminar a laminação

das chapas acima de Ar_3 , quando o esboço ainda se encontra totalmente austenítico. Mas, no caso de alguns produtos, é até recomendável que a temperatura final de laminação fique um pouco abaixo de Ar_3 para se conseguir uma melhor combinação de resistência mecânica e tenacidade no produto final, através da laminação de fase predominantemente austenítica com pequena fração de ferrita (Mayo et al., 2020). Gemini, na correção dessa resposta, concordou com CoPilot, ao passo que Claude e DeepSeek assumiram uma posição ortodoxa e criticaram essa recomendação do CoPilot, apesar dela ser mais compatível com a realidade industrial.

- Na resposta à questão 5 (principais diferenças de processo entre a laminação de chapas grossas e a laminação de tiras a quente), DeepSeek afirmou que, quanto maior a taxa de deformação, menor a temperatura em que a recrystalização ocorre. Contudo, isso está em desacordo com o parâmetro de Zener-Hollomon:

$$Z = \dot{\epsilon} e^{\frac{Q}{RT}}$$

onde $\dot{\epsilon}$ é a taxa de deformação, Q a energia de ativação, R é a constante universal dos gases e T é a temperatura. Quanto maior for o valor de Z , mais difícil é a recrystalização. Portanto, de acordo com a fórmula, quanto maior for a taxa de deformação, também a temperatura deverá ser mais alta para que ocorra recrystalização (Siciliano Jr; Jonas, 2000). Curiosamente, DeepSeek citou essa fórmula, mas foi incapaz de interpretá-la matematicamente.

- Nas respostas à questão 6 (grau mínimo de deformação para ocorrência de recrystalização completa da austenita), ChatGPT, CoPilot, Gemini e DeepSeek inicialmente informaram que o grau mínimo de deformação para ocorrência de recrystalização da austenita durante a laminação a quente estaria na faixa entre 0,10 e 0,15. Na realidade, são valores baixos demais. Eles até podem promover o início da recrystalização, que ocorre em pontos favoráveis da microestrutura, como contornos de grão, mas ela sofre interrupção antes de sua conclusão devido à energia insuficiente proporcionada pelo encruamento decorrente da baixa deformação. Mas esses LLMs entraram em contradição posteriormente, afirmando que, na laminação industrial, essa faixa na verdade estaria entre 0,15 e 0,20, que é correta. Curiosamente, apesar de terem cometido o mesmo erro em suas respostas, no momento das correções de outros LLMs essa incoerência foi apontada pelo CoPilot e DeepSeek. Aparentemente houve um conflito entre as informações

disponíveis nos LLMs que levou a um raciocínio incoerente. Além disso, verificou-se ainda os LLMs não conseguiram distinguir adequadamente as diferenças entre as recristalizações dinâmica (que ocorre simultaneamente com a deformação) e estática (que ocorre após a deformação) (Siciliano Jr; Jonas, 2000). E Chat GPT tornou a cometer a mesma confusão entre RLT, RST e T_{nr} já descrita na resposta à questão 2.

- Na resposta à questão 7 (deformação mínima para ocorrência de recristalização dinâmica), ChatGPT afirmou, erradamente, que um tamanho de grão mais grosseiro na microestrutura reduz a deformação necessária para a ocorrência da recristalização dinâmica. CoPilot concordou com essa afirmação errônea em sua correção. Além disso, ChatGPT não informou a fórmula do parâmetro Z de Zener-Hollomon (já citada num comentário sobre a resposta à questão 5), que define as condições da ocorrência de recristalização (Siciliano Jr; Jonas, 2000). Claude e DeepSeek apontaram esses erros.
- Na resposta a essa mesma pergunta, DeepSeek afirmou que a recristalização dinâmica da austenita seria benéfica para o refino microestrutural durante a fase de esboçamento da laminação controlada. Mas o fato é que, na prática industrial, as condições metalúrgicas nessa fase do processo não favorecem a ocorrência da recristalização dinâmica (Siciliano Jr; Jonas, 2000). ChatGPT concordou que a ocorrência de recristalização dinâmica no esboçamento seria benéfica, mas acertadamente ponderou que ela seria muito improvável sob condições industriais.
- Ao corrigir a resposta da questão 10 (diferença entre laminação de normalização e laminação controlada) elaborada pelo Gemini, o DeepSeek alucinou, afirmando que a designação “laminação de normalização” vale inclusive para chapas grossas que são posteriormente submetidas a um tratamento térmico adicional de normalização. Isso é totalmente errado, pois o processo de laminação de normalização foi concebido justamente para suprimir esse tratamento posterior (Matsubara et al., 2025). DeepSeek repetiu esse erro na sua resposta a essa mesma pergunta, o que foi notado pelo ChatGPT e Claude. Gemini também alucinou, tendo concordado com essa afirmação que estava presente na resposta do DeepSeek à essa questão.
- Curiosamente, o ChatGPT espontaneamente declarou, ao final das correções das respostas do questionário elaboradas pelo CoPilot e Claude, que elas tinham grande chance

de terem sido geradas por inteligência artificial, em função de seu estilo didático e divisão num número excessivo de itens, bem como seu grande tamanho. CoPilot fez esse mesmo comentário sobre as respostas de DeepSeek. Segundo os LLMs avaliadores, essa constatação, se confirmada, constituiria um demérito para candidatos humanos submetidos a uma avaliação profissional.

5 CONCLUSÃO

Este trabalho mostrou, de forma algo surpreendente, que *chatbots* baseados em LLMs apresentaram um domínio relativamente satisfatório sobre o tema da laminação controlada de aços estruturais microligados, mesmo em sua versão pública e gratuita. Para aqueles que, há algumas poucas décadas atrás, tinham grande dificuldade em obter fontes de conhecimento sobre esse assunto em livros e artigos técnicos com acesso restrito, é gratificante verificar que, hoje, estudantes, técnicos e engenheiros em início de carreira podem obter informação relativamente boa sobre esse assunto, gratuitamente, e em qualquer lugar que disponha de conexão à internet, desde que devidamente orientados sobre o uso correto dessas plataformas e, particularmente, para sempre manter uma visão crítica sobre suas respostas.

A avaliação global do conhecimento das LLMs na área específica estudada nesse trabalho não variou muito, com notas variando desde 4,3 (para um máximo de 5,0) para o CoPilot até 4,8 para o Claude. As demais plataformas, ChatGPT, Gemini e DeepSeek, ficaram no grupo intermediário. Gemini, com nota de 4,5, conseguiu o segundo lugar e constitui uma boa opção para quem quer obter conhecimento de qualidade razoável, mas de forma concisa, sem ter de enfrentar o excessivo detalhamento que caracterizou as respostas do Claude.

Ainda assim, foram observadas várias incorreções, algumas até graves, sobre os conceitos metalúrgicos envolvidos no tema aqui estudado, tanto nas respostas do questionário como em suas correções. Constatou-se às vezes incoerências ao longo do tempo, ou seja, um mesmo conceito foi apresentado de forma certa ou errada conforme a situação envolvida. Portanto, deve-se frisar novamente que as respostas geradas pelos LLMs não são determinísticas, mas possuem natureza estatística e estão sujeitas a flutuações de tamanho e quantidade conforme o contexto envolvido. Além disso, essas respostas podem ter sua veracidade afetada por “alucinações”, que podem ocorrer pelos mais diversos motivos: resposta do LLM através de geração de texto plausível, mas não necessariamente verdadeiro; o viés de aprovação; desconhecimento do assunto a que se refere a consulta, entre outros.

A abordagem aqui adotada, baseada na automatização da geração do conhecimento por LLMs e sua posterior correção usando o conceito de *LLM-as-a-Judge*, revelou-se útil para determinar qual LLM comercial e gratuito é o melhor para tratar do tópico específico da laminação controlada de aços estruturais microligados. Contudo, essa abordagem não é a mais indicada para reduzir o risco de “alucinações”. A engenharia de *prompts* oferece várias possibilidades que permitem minimizar esse tipo de problema. São obtidos resultados com qualidade muito

melhor e menor incerteza utilizando-se a técnica do *prompting* interativo, ou seja, por meio do desenvolvimento de uma conversa progressiva e crítica entre o usuário e LLM (*Human-in-the-Loop*), o que promove o surgimento de *insights* muito úteis, ainda que eles tenham de ser posteriormente confirmados em fontes confiáveis. Outra opção de *prompt* é o chamado *chain of thought*, onde o LLM tem de apresentar seu raciocínio passo a passo antes de elaborar sua conclusão. É interessante notar que o CoPilot e DeepSeek externaram automaticamente essa abordagem em algumas de suas respostas e correções, sem que isso lhes tenha sido solicitado. Dessa forma, os LLMs podem atuar nos papéis em que eles desempenham melhor, ou seja, na organização do raciocínio e compilação do conhecimento. Frequentemente o caminho mais interativo também é o mais proveitoso: a jornada pode ser muito melhor que o próprio destino.

REFERÊNCIAS

- ABBASI, M. et al. Comparative Performance Analysis of Large Language Models for Structured Data Processing: An Evaluation Framework Applied to Bibliometric Analysis. **Applied Sciences**, vol. 16, 669, 30 p., 2026. Disponível em <https://www.mdpi.com/2076-3417/16/2/669>. Acesso em: 01 mar. 2026.
- ANTHROPIC. **The Claude 3 Model Family: Opus, Sonnet, Haiku**. Disponível em https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Mo-del_Card_Claude_3.pdf. 2024. San Francisco, E.U.A.: Anthropic, PBC. Acesso em 29. mar 2026.
- BALLAGER, A. et al. **RAG vs fine-tuning: pipelines, tradeoffs, and a case study on agriculture**. [arXiv:2401.08406v3](https://arxiv.org/abs/2401.08406v3) [cs.CL], 2024. Disponível em <https://doi.org/10.48550/arXiv.2401.08406>. Acesso em: 28 feb. 2026.
- BERNETIC, J. et al. Centreline Formation of Nb(C,N) Eutectic in Structural Steel. **Metalurgija**, vol.49, n.1, p. 29-32, 2010. Disponível em <https://hrcak.srce.hr/40246>. Acesso em 24 maio 2026.
- BHAVSAR, P. **Mastering LLM-as-a-Judge – Explore How to Evaluate AI Outputs with High Precision Using LLMs**. Burlingame, E.U.A.: Galileo AI Inc., 2025. Disponível em <https://galileo.ai/mastering-llm-as-a-judge>. Acesso em: 01 mar. 2026.
- BHAVSAR, P. **Mastering RAG - A Comprehensive Guide for Building Enterprise-Grade RAG Systems**. Burlingame, E.U.A.: Galileo AI Inc., 2024. Disponível em <https://galileo.ai/mastering-rag>. Acesso em: 02 maio 2026.
- BOLOTA, J.R. (editor). **Siderurgia, uma Visão Geral**. São Paulo, Brasil: Associação Brasileira de Metalurgia, Materiais e Mineração, 2024, 570 p.
- CHEN, B. et al. Unleashing the potential of prompt engineering for large language models. **Patterns**, vol.6, n.6, 44 p., June 13, 2025. Disponível em <https://doi.org/10.1016/j.pat-ter.2025.101260>. Acesso em: 05 abr 2026/
- DEEPSEEK-AI. **DeepSeek-V3 Technical Report**. [arXiv:2412.19437v2](https://arxiv.org/abs/2412.19437v2) [cs.CL], 2025. Disponível em <https://doi.org/10.48550/arXiv.2412.19437>. Acesso em 15 mar. 2026.
- GEMINI TEAM. **Gemini: A Family of Highly Capable Multimodal Models**. [arXiv:2312.11805](https://arxiv.org/abs/2312.11805) [cs.CL], 2025. Disponível em <https://doi.org/10.48550/arXiv.2312.11805>. Acesso em 29 mar. 2026.
- GUIMARÃES, E. et al. **Analyzing Prominent LLMs: An Empirical Study of Performance and Complexity in Solving LeetCode Problems**. In: EASE '25: 29th INTERNATIONAL CONFERENCE ON EVALUATION AND ASSESSMENT IN SOFTWARE ENGINEERING, 2025, Istanbul. **Proceedings [...]**. New York: ACM, 2015. P.949-958. Disponível em <https://dl.acm.org/doi/epdf/10.1145/3756681.3756983>. Acesso em 15 mar. 2026.

HAMRYSZCZAK, T. Hot Rolling of HSLA Steels - A Review of Recent Studies. **Archives of Metallurgy and Materials**, v.69, n.3, p. 1151-1157, 2024. Disponível em https://www.imim.pl/files/archiwum/Vol3_2024/38.pdf. Acesso em: 28 feb. 2026.

HOSKING, T.; BLUNSOM, P.; BARTOLO, M. **Human Feedback is Not Gold Standard**. [arXiv:2309.16349](https://doi.org/10.48550/arXiv.2309.16349) [cs.CL], 2024. Disponível em <https://doi.org/10.48550/arXiv.2309.16349>. Acesso em: 19 feb. 2026.

INSTITUTO AÇO BRASIL. **Anuário Estatístico 2025** = *Brazil Steel Databook 2025*. Rio de Janeiro: Instituto Aço Brasil, ago. 2025. Anual. ISSN 1806-3195. Disponível em: https://www.acobrasil.org.br/site/wp-content/uploads/2025/08/AcoBrasil_Anuario_2025-1.pdf. Acesso em: 7 jun. 2026.

IPEIROTIS, P.G.; PROVOST, F.; WANG, J. Quality management on Amazon Mechanical Turk. *In: KDD '10: THE 16th ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING*, 2010, Washington DC. **Proceedings [...]**. New York: ACM, 2010. P.64-67. Disponível em <https://dl.acm.org/doi/10.1145/1837885.1837906>. Acesso em: 08 mar 2026.

KLESEL, M.; WITTMANN, H.F. Retrieval-Augmented Generation (RAG). **Business & Information Systems Engineering**, vol. 67, n.4, p. 551–561, 2025. Disponível em <https://link.springer.com/article/10.1007/s12599-025-00945-3>. Acesso em: 28 feb. 2026.

LEYSEN, F. et al. Influence of Coiling Temperature on Toughness of Hot Rolled Microalloyed Steels. *Materials Science Forum*, v. 284-286, p. 263-70, 1998. Disponível em <https://10.4028/www.scientific.net/MSF.284-286.263>. Acesso em 24 maio 2026.

LIU, P. et al. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. **ACM Computing Surveys**, vol. 55, n. 9, Article 195, 35 p., 16 January 2023. Disponível em <https://dl.acm.org/doi/10.1145/3560815>. Acesso em: 01 mar. 2026.

LOPEZ, B. et al. The Non-Recrystallization Temperature in Nb-based Steels: from Laboratory to Industrial Conditions. *In: 6th INTERNATIONAL CONFERENCE ON THERMOMECHANICAL PROCESSING*, 2022, Shenyang (China). **Proceedings [...]**. P.249-254.

MATSUBARA, D.B. et al. Normalizing Rolling of Niobium Microalloyed Heavy-Gauge Plates for Wind Tower Applications. *In: 13th INTERNATIONAL ROLLING CONFERENCE*, 2025, Dallas. **Proceedings [...]**. Pittsburgh: AIST, 2025, P. 86-93

MAYO, U. et al. Relation Between Microstructure and Mechanical Properties on Intercritically Deformed Low Carbon Steels. **Procedia Manufacturing**, v. 50, p. 291–297, 2020. Disponível em <https://doi.org/10.1016/j.promfg.2020.08.054>. Acesso em 24 maio 2026.

MOHRBACHER, H. Application of Microalloying for Controlling Recrystallization and Grain Growth During Downstream Steel Processing. **Metallurgical and Materials Transactions**, vol. 57A, n. 6, p. 2548-2571, June 2026. Disponível em <https://doi.org/10.1007/s11661-026-08172-5>. Acesso em 12 maio 2026.

OKUMURA, N. et al. Relation between Hot Rolling Conditions and Mechanical Properties

of Heavy Plates Produced from Thin Continuously-Cast Slabs. **Tetsu-to-Hagané**, vol. 70, n. 3, p. 388-395, 1984. Disponível em https://doi.org/10.2355/tetsutohagane1955.70.3_388. Acesso em 23 maio 2026.

OUYANG, L. et al. Training Language Models to Follow Instructions with Human Feedback. 2022. [arXiv:2203.02155v1](https://arxiv.org/abs/2203.02155) [cs.CL]. Disponível em <https://doi.org/10.48550/arXiv.2203.02155>. Acesso em 24 maio 2026.

OPEN AI et al. **GPT-4 Technical Report**. 2024. [arXiv:2303.08774v6](https://arxiv.org/abs/2303.08774) [cs.CL], 2024. Disponível em <https://doi.org/10.48550/arXiv.2303.08774>. Acesso em: 15 feb. 2026.

SAHOO, P. et al. **A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications**. 2025. [arXiv:2402.07927v2](https://arxiv.org/abs/2402.07927) [cs.AI]. Disponível em <https://doi.org/10.48550/arXiv.2402.07927>. Acesso em 05 abr. 2026.

SCHNEIDER, J.; MESKE, C.; KUSS, P. Foundation Models A New Paradigm for Artificial Intelligence. **Business & Information Systems Engineering**, vol. 66, n.2, p. 221–231, 2024. Disponível em <https://dx.doi.org/10.1007/s12599-024-00851-0>. Acesso em: 01 mar. 2026.

SHI, S. et al. **Judging the Judges: A Systematic Study of Position Bias in LLM-as-a-Judge**. [arXiv:2406.07791v9](https://arxiv.org/abs/2406.07791) [cs.CL], 2025. Disponível em <https://doi.org/10.48550/arXiv.2406.07791>. Acesso em: 30 mar. 2026.

SICILIANO JR, F.; JONAS, J.J. Mathematical Modeling of the Hot Strip Rolling of Microalloyed Nb, Multiply-Alloyed Cr-Mo, And Plain C-Mn Steels. **Metallurgical and Materials Transactions A**, vol. 31A, n. 2, p. 511-530, 2000. Disponível em <https://doi.org/10.1007/s11661-000-0287-8>. Acesso em: 23 maio 2026.

STALHEIM, D.G.; GLODOWSKI, R. Production of Fine Grained As-Rolled Structural Plate Steels. In: AISTECH 2010 - IRON & STEEL TECHNOLOGY CONFERENCE, 2010, AIST, Pittsburgh PA. **Proceedings [...]**. Pittsburgh, AIST, 2010. Vol. II, P.1019-1025.

STREISSELBERGER, A.; SCHWINN, V.; HUBO, R. Microalloyed Structural Plate Rolling, Heat Treatment and Applications. In: NIOBIUM 2001 – SCIENCE & TECHNOLOGY, 2001, TMS, Orlando FL. **Proceedings [...]**. Pittsburgh, TMS, 2001. P.625-646

TIAN, S. et al. Steel design based on a large language model. **Acta Materialia**, vol. 285, 120663, 13 p., 15 February 2025. Disponível em <https://doi.org/10.1016/j.actamat.2024.120663>. Acesso em: 6 jan. 2026.

VILLALOBOS, J.C. et al. Microalloyed Steels Through History until 2018: Review of Chemical Composition, Processing and Hydrogen Service. **Metals**, vol. 8, 351, 49 p., 2018. Disponível em <https://doi.org/10.3390/met8050351>. Acesso em 04 abr. 2026.

WANG, P. et al. **Large Language Models are not Fair Evaluators**. [arXiv:2305.17926](https://arxiv.org/abs/2305.17926) [cs.CL], 2023. Disponível em <https://doi.org/10.48550/arXiv.2305.17926>. Acesso em 27 mar. 2026.

ZHAO, W.X. et al. **A Survey of Large Language Models**. [arXiv:2303.18223](https://arxiv.org/abs/2303.18223) [cs.CL], 2026. Disponível em <https://doi.org/10.48550/arXiv.2303.18223>. Acesso em 21 mar 2026.

APÊNDICE A – Questionário Apresentado aos LLMs

“Preamble and Motivation:

You are a junior engineer applying for a supervisory position in the hot flat steel rolling line. One of the prerequisites for your approval will be to answer a ten-question questionnaire on the metallurgy and practice of controlled rolling (Thermomechanical Controlled Processing - TMCP) of flat structural microalloyed steels, to verify your level of knowledge on the subject and ascertain whether you possess an adequate level of technical expertise to assume this position.

Please focus on reliable and trustworthy sources.

The questions will be presented now, one at a time.

Good luck!”

Questions:

- 1. How to determine the recommended slab reheating temperature for structural microalloyed steel plates to be rolled using controlled rolling? Please consider all contradictory metallurgical, process and cost objectives.*
- 2. How to define an optimized distribution of total strain (slab-plate or slab-strip) between the roughing and finishing phases during controlled rolling of microalloyed steels, aiming to have an optimum balance between mechanical strength and toughness of flat products made with microalloyed steels? Remember that the productivity of the rolling mill also is a very important aspect.*
- 3. What are the recommended conditions for the pass schedule to be applied during the roughing phase of controlled rolling of Nb microalloyed steels in a plate mill, aiming at a balance between mechanical properties of the final product and mill productivity?*
- 4. What are the recommended conditions for the pass schedule to be applied during the finishing phase of controlled rolling of Nb flat microalloyed steels in a plate mill, aiming at a balance between mechanical properties of the final product and mill productivity?*
- 5. What are the main process and metallurgical differences between controlled rolling of flat microalloyed steels performed in a heavy plate mill and in a hot strip mill?*
- 6. What is the minimum strain degree that must be applied to work-hardened austenite for it to fully recrystallize, assuming it is above the RLT (Recrystallization Limit Temperature)?*

7. *What process and metallurgical parameters influence the determination of the critical strain above which dynamic recrystallization of austenite occurs? Is this metallurgical phenomenon positive for the final properties of the product?*
8. *How do you calculate the value of the average grain size when partial recrystallization of austenite occurs between hot rolling passes? Please explain the assumptions and details of the calculation.*
9. *What are the effects of different microalloying elements (niobium, vanadium, titanium) during the thermomechanical treatment and on the final mechanical properties of flat hot rolled structural products submitted to controlled rolling? Which factors justify the adoption of a specific alloy design with different microalloying elements?*
10. *What is the difference between the thermomechanical treatments applied to microalloyed structural flat steels, known as "normalizing rolling" and "controlled rolling"? What leads to the choice of one of them?*

APÊNDICE B – Procedimento para Avaliação das Respostas dos LLMs ao Questionário

“Motivation:

You are a metallurgical engineer and experienced professional in the field of hot rolling of flat products made of microalloyed structural steels and a specialist in controlled rolling (Thermomechanical Processing Controlled Processing - TMCP), with a master's and doctorate in this area. You will now evaluate the answers to ten questions on this subject, submitted by a junior engineer who is applying for a supervisory position in your department, in order to verify if he possesses the appropriate technical background for the role. You should evaluate the answers according to the following items, providing a score from 1 to 5 for each. Assess each criterion independently:

- *Technical correctness, accuracy. How factually accurate and correct is the output relative to ground truth or verifiable information?*
- *Scope, relevance. How well the output addresses the question?*
- *Metallurgical coherence.*
- *Reasoning Coherence. Consider logical flow and internal consistency of the thought process.*
- *Logical Consistency. Absence of contradictions within a single response or across multiple responses.*
- *Correct use of technical terminology.*
- *Clarity, depth, organization, and comprehensibility of explanations provided.*
- *Objectivity, concision.*
- *Completeness. Does the response cover all aspects of the query without omissions? But note that the quality of the answer is more important than its length.*
- *Hallucination rate.*

Then please calculate the overall average note for this question from the items above.

The questions will now be presented, one at a time.

Thank you for your cooperation!

After the evaluation of the last answer:

The final question: do you consider the candidate technically suitable for this position?”

APÊNDICE C

Declaração de uso de INTELIGÊNCIA ARTIFICIAL, cuidados éticos e adequação à LGPD no TCC

MBA em Inteligência Artificial e Big Data – ICMC/USP

Aluno(a): Antonio Augusto Gorni

Título do TCC: Avaliação do Conhecimento dos Grandes Modelos de Linguagem Públicos sobre a Laminação Controlada de Aços Microligados

Declaro que, na elaboração deste Trabalho de Conclusão de Curso:

☐ Não utilizei ferramentas de Inteligência Artificial Generativa como apoio à elaboração do TCC.

☒ Utilizei ferramentas de Inteligência Artificial Generativa como apoio à elaboração do TCC, conforme indicado abaixo.

1. Como utilizei IA

Marque as opções aplicáveis:

☐ Revisão, aprimoramento ou tradução de textos

☒ Busca, síntese ou organização de informações e literatura

☐ Geração, revisão ou depuração de código

☒ Análise, tratamento ou interpretação de dados

☐ Apoio à estruturação de ideias, metodologia ou solução

☐ Geração ou aprimoramento de imagens, gráficos ou outros conteúdos

☒ Outro: Avaliação do conhecimento de chatbots públicos sobre um tema específico

Ferramentas utilizadas: ChatGPT, CoPilot, Gemini, Claude, DeepSeek

2. Nível de utilização

Marque a opção que melhor representa o uso de IA na elaboração do TCC:

☐ **Nível 1 – Apoio pontual**

Uso em atividades específicas, como revisão de texto, tradução, esclarecimento de dúvidas ou sugestões pontuais.

☒ **Nível 2 – Apoio significativo**

Uso recorrente em algumas etapas, como estruturação de conteúdo, programação, análise de dados ou desenvolvimento da solução.

☐ **Nível 3 – Apoio substancial**

Uso relevante na geração ou desenvolvimento de conteúdos, códigos, análises ou outros elementos que integram o TCC.

3. Declaração sobre cuidados Éticos e atendimento à LGPD

Cuidados éticos: quais foram os procedimentos éticos adotados em pesquisas envolvendo seres humanos, conforme descrito no formulário original.

☒ Não se aplica

☐ Cuidados adotados:

Quando o estudo envolver coleta, uso ou análise de dados pessoais, indicar as medidas adotadas para proteção de dados pessoais em conformidade com a LGPD.

x Não se aplica

☐ Medidas adotadas:

4. Declaração de responsabilidade

Declaro que revisei, validei criticamente e assumo a responsabilidade pelos conteúdos produzidos com o auxílio de ferramentas de Inteligência Artificial.

Declaro, ainda, que compreendo plenamente, sou capaz de explicar e defender todo o conteúdo apresentado neste TCC, bem como que observei os princípios éticos e os requisitos de proteção de dados aplicáveis.

Data: 19/08/2026

Assinatura do(a) aluno(a):

