ASSESSMENT OF THE KNOWLEDGE OF PUBLIC LARGE LANGUAGE MODELS ABOUT THERMOMECHANICAL CONTROLLED PROCESSING OF MICROALLOYED STEELS*

Antonio Augusto Gorni[1]
Fernando Pereira dos Santos[2]

Abstract
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, 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 detail levels among the LLMs assessed. 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.
Keywords: Artificial Intelligence; Large Language Models; Microalloyed Steels; Controlled Rolling.

[1] *PhD, MSc., Materials Engineer, Independent Metallurgical Consultant, São Vicente SP, Brazil.*
[2] *PhD, MSc, Computer Science, Artificial Intelligence Specialist, Venturus - Innovation & Technology, Campinas, SP, Brazil.*

1 INTRODUCTION

The digital revolution began half a century ago, when the price of digital circuits plummeted, enabling the popularization of microcomputers with processing power that, until then, was the privilege of large companies or institutions. Less than twenty years later, the spread of the Internet signified another revolution, allowing communication between computers at extremely low costs and a massive and instantaneous exchange of information. The uninterrupted progress in these two areas has now made the use of artificial intelligence tools viable at virtually no cost for users.

The first language models, designed to guess the next word in a text, when equipped with transformer neural networks, with complex architecture, and subjected to special learning techniques using the enormous amount of data available on the internet, suddenly acquired capabilities unexpected even for their own creators, such as text generation, translation, document summarization, computer programming, and answering questions, including a sophisticated conversational capability. At the end of 2022, ChatGPT emerged, the first chatbot based on Large Language Models (LLM) made available to the public, even for free. Then, in 2023, CoPilot, Claude, and Gemini appeared, and in 2025, DeepSeek.

However, this progress does not occur without some setbacks. One of them is the enormous computational resources needed to train these LLMs, at a prohibitive cost that can only be borne by large corporations, making independent academic participation in this process unfeasible. Because of this, it is not known exactly what data sources were used to train these models. This constitutes a serious problem for the reliability of LLMs because, when exposed to questions about knowledge in which they were not trained, they tend to “hallucinate," i.e., provide seemingly logical responses that are completely detached from reality. Worse still, the high conversational capacity of LLMs generates extremely convincing hallucinations that can deceive the most unwary. And one must remember that LLM models do not seek truth; but rather the most plausible text given the context. Truth and plausibility often coincide, but not always. Besides that, LLMs are trained with human feedback — and humans tend to positively evaluate responses that agree with their assumptions. This creates a structural tendency to validate what the user says, even when it is wrong.

The objective of this study was to verify the possibility of using LLMs to answer technical questions related to the controlled rolling of structural microalloyed steels. Since this is a highly specific topic with significant professional responsibility, the reliability of LLMs is of crucial importance – i.e., it is necessary to minimize the risk of hallucinations while the model must provide relevant and complete information.

There are several ways to minimize the problem of hallucinations. One of them is so-called fine-tuning, which consists of updating an LLM using additional documents relevant to a specific topic. In this way, LLM's performance improves when dealing with this subject, eliminating the enormous effort and financial investment required for a foundational model. Another possibility is the creation of a separate database with vectorized information, i.e., the so-called Retrieval Augmented Generation (RAG), which will be consulted by the LLM. This alternative requires even fewer resources than fine-tuning an LLM, but its implementation is also not simple and requires powerful servers [1]. On the other hand, the use of so-called prompt engineering, i.e., a careful elaboration of the dialogue to be established with the LLMs, can reduce the chance of hallucinations without significant costs [2]. For example, the use of the so-called "role prompt" – i.e., the role that the LLM should assume when responding to a query. This helps the LLM to adapt to the task at hand, adjusting its style, reasoning context,

technical focus, and level of detail, among other aspects. [3]. Table 1 summarizes the characteristics of these three approaches [4].

Currently, it is very tempting to use the resources of free LLMs to resolve doubts or obtain specialized technical knowledge, tasks that can take hours or even days using conventional knowledge sources. This raises the question: among the various free public LLM options available today, which would be the best for a given topic? We cannot know beforehand, since the knowledge sources used in their training are not disclosed.

Table 1: Comparation between Fine Tuning, RAG and Prompt Engineering [4].

Characteristic	Fine Tuning	RAG	Prompt Engineering
Can it make use of external knowledge sources?	x	√	x
Does it minimize hallucinations?	√	√	√
Does it require domain-specific data?	√	x	x
Is it suitable for dynamic data?	x	√	x
Does it offer clear interpretability of outputs?	x	√	x
Low resource utilization	x	x	√
Quick deployment	x	x	√

Evaluating LLMs is a topic that has become increasingly important, as it is fundamental to their improvement and increased reliability. However, it is not an easy task, due to their non-deterministic nature – i.e., their responses to the same question can vary over time depending on their statistical character and, in the case of free versions, are probably more sensitive to the level of demand they are experiencing at any given time. The first proposal involves having LLMs evaluated by humans. However, the results tend to be inconsistent due to personal biases, unless there is a properly trained body of evaluators [5]. Furthermore, the amount of work involved makes it necessary to remunerate the evaluators, which restricts the use of this resource in academic activities.

For this reason, the “LLM-as-a-judge” approach emerged, where LLMs themselves evaluate other LLMs. This is particularly suitable when the metrics for evaluating model responses are not numerical. Other advantages are the speed and low costs inherent in this approach, in addition to minimizing the biases typical of human evaluations. But LLMs also exhibit some biases when correcting responses. For example, responses with better aesthetics or greater length tend to be better evaluated than more complete responses but written in a simpler way. When presented with a sequence of responses to the same question, the first ones are better evaluated than the others, even if the latter are of better quality [6].

Based on everything discussed, this study was developed with the objective of evaluating the performance, in terms of knowledge and reliability, of several publicly available and free LLMs within the specific theme of controlled rolling of structural microalloyed steels. As this is a very specific topic with high professional responsibility, the reliability of LLMs assumes crucial importance – i.e., it is necessary to minimize the risk of hallucinations, while at the same time the models must provide relevant and complete information. The controlled rolling of microalloyed steels has motivated the publication of hundreds, perhaps thousands, of technical papers in the last sixty years, and it is quite likely that a portion of this body of work has already been used in the learning of LLMs.

2 DEVELOPMENT

2.1 Methodology

Five publicly available, free LLMs were chosen for this evaluation: ChatGPT (https://chatgpt.com), CoPilot (https://copilot.microsoft.com), Gemini (https://gemini.google.com), Claude (https://claude.ai), and DeepSeek (https://www.deepseek.com). The aim was to use them in the most accessible way, so that the evaluation could be useful for students, technicians, and engineers at the beginning of their careers.
The evaluation of the LLMs was defined as follows: each of them would response questions from a test on the controlled rolling of structural microalloyed steels, with the responses being corrected by the remaining four. The use of the English language was defined for the tests and corrections, since most of the knowledge on the subject is written in that language, thus avoiding the disturbances resulting from translations between Portuguese and English, especially the jargon typical of this topic.
The "role prompt" approach was adopted, both for administering the test and for grading the responses. The test began by creating a specific persona, that is, a candidate for a supervisory position on a hot rolling mill line who would have to undergo an examination to verify if they had the appropriate level of technical knowledge for the job. The following text was used in the test's "role prompt":

"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 response 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!"

The following questions were then presented to the candidate, one at a time, and the corresponding response was obtained; both were then immediately transcribed into a Word file:

"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 is also a particularly 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 mean 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?"

Once all the responses were obtained, the correction procedure began with a new "role prompt", now creating a professor-doctor in charge of this task, as shown below:

"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 responses to ten questions on this subject submitted by a junior engineer who is applying for a supervisory position in your department, to verify if he possesses the appropriate technical background for the role. You should evaluate the responses 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 does 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 response is more important than its length.*
- *Hallucination rate.*

Then please calculate the overall mean note for this question from the items above.
The questions will now be presented, one at a time.
Thank you for your cooperation!"

Next, the question statement was presented, one by one, along with the candidate's corresponding response. Presenting responses to different questions sequentially avoids the bias typical of LLMs when they analyze different responses to the same question, i.e., decreasing scores as they analyze the sequence of responses. The corrections obtained for each question were transcribed into a Word file. Finally, after the last question had been corrected, the final question was posed to the evaluator:

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

After grading all candidates' exams, the author analyzed the texts of the LLMs' responses and their respective corrections, as well as the tabulation and analysis of the scores. The selection of the best LLM was based on the mean of these scores assigned by the others.
Efforts were made to administer the exams and their respective grading within the shortest possible timeframe, to maintain stable performance among the LLMs, both in exam administration and grading. Furthermore, before administering or grading an exam, the LLM was asked to indicate the current version of the exam and the date of its last update. Attempts were made to obtain the temperature parameter value at that moment, but without success, since this information is not provided to users on any of the platforms analyzed here. This prevents a more detailed analysis of the reproducibility of the responses and corrections obtained.

2.2 Results and Discussion

2.2.1. Responses to the Questionnaire

The five tests were administered on April 9th, 2026. At the time they prepared the responses to the questions, the LLMs were operating under the following versions: ChatGPT, model GPT-5, current version GPT-5.2, last training update, August 2025; Gemini, model Gemini 3 Flash, continuously updated; Claude, model Sonnet 4.5, last update at the end of January 2025; and DeepSeek, version DeepSeek-V3, variant DeepSeek-V3-0324, last update on March 24, 2025. CoPilot, in turn, reported that it did not have a traditional model or version number like the other platforms, but that it was continuously updated.
The responses were obtained quickly and uninterruptedly for all LLMs, except for Claude, who interrupted the free service after the fourth question and only resumed activities about four hours later. Most of the time the responses did not include graphs, except in a few cases; but they were extremely simple, generated with characters.

Table 2 shows the number of characters in each response generated by the LLMs to the test questions. Overall, it can be observed that Gemini generated responses with minimal size and relative standard deviation, exactly the opposite of Claude. ChatGPT and CoPilot also generated responses with sizes below the observed mean, while the opposite occurred with DeepSeek. This reflects the greater detail of the responses elaborated by DeepSeek and Claude compared to the other LLMs.

Table 3 shows, for each LLM, the total number of characters, time spent on preparation, speed, and number of pages of the test responses. The execution time is approximate, as it also included the manual transcription of the responses generated by the LLMs into a Word file. The number of pages of the responses is excessively high; the minimum number was 23 for Gemini and reached 58 pages for Claude. This clearly shows that the tests were performed by LLMs, as these are unrealistic numbers for human candidates, limited by the time and knowledge available. The five sets of responses to the questionnaire totaled 204 pages, which, in practice, makes their correction by humans unfeasible, unless they have the necessary time and interest to conduct this analysis.

Table 2: Number of characters (without spaces) of the responses to the questions by the candidate LLMs. R.S.D. means Relative Standard Deviation.

Questions	1	2	3	4	5	6	7	8	9	10	Mean	R.S.D. [%]
Chat GPT	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%
Mean	3194	4230	5332	6710	4929	3624	5040	4348	7402	7155	5196	28%
R.S.D. [%]	16%	23%	47%	61%	31%	47%	46%	54%	57%	64%		

Table 3: Total number of characters (without spaces), time spent, speed and number of pages of the responses to the questions by the candidate LLMs.

Candidate	Characters	Time Spent [min]	Speed [characters/min]	Pages
Chat GPT	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
Mean	51964	23	330	Total
R.S.D. [%]	45%	26%	63%	204

2.2.2. Corrections of the Responses

The corrections of the five sets of ten responses each by four LLMs were made between April 11th and 15th, 2026. The data regarding the LLM models and versions were the same as described in the previous section. Here too, only Claude interrupted the process after the fourth correction, resuming activities about four hours later, creating some disruption as it was necessary to inform him again about the initial prompt, as well as the evaluations of the questions that had been made previously.

ChatGPT and CoPilot, however, "got lost" in the middle of the correction process, adopting a correction procedure different from what had been agreed upon in the initial prompt. ChatGPT even started expressing himself in Portuguese at that moment. There was only one such occurrence for each of these two LLMs, requiring the procedure to be interrupted and the initial prompt re-presented so that the correction could proceed as agreed.

As shown in Table 4, the character count of the corrections showed the same pattern observed in the responses to the questions, except for CoPilot, which now lagged behind ChatGPT. Once again, Claude provided the most detailed corrections, while Gemini generated the shortest corrections in the set. Table 5 shows that the number of pages in the corrections, including the question statements and their respective responses, was in the order of several hundred for each LLM, reaching a total of almost 1,500 pages. This reflects, to varying degrees, the LLMs' priority in maximizing the completeness of the responses and corrections, often far exceeding what was requested in the question statements. Despite the larger size of its corrections, Claude was the fastest LLM, followed by DeepSeek, ChatGPT, Gemini, and CoPilot.

Table 4: Total number of characters (without spaces) of the corrections to the responses of the questions by the evaluator LLMs.

Evaluator	1	2	3	4	5	6	7	8	9	10	Mean	R.S.D. [%]
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	9405	10718	10836	11670	11419	11954	13399	11330	11454	12392	11458	9%
Claude	14108	16695	17768	19235	16835	18350	19057	18991	18968	19945	17995	10%
DeepSeek	12199	16779	17934	18901	18707	19031	17406	17277	17477	17220	17293	11%
Mean	12526	14922	15577	16643	16009	16516	17335	16931	17154	16483		
R.S.D. [%]	17%	18%	20%	18%	18%	18%	16%	23%	23%	17%		

Table 5: Total number of characters (without spaces), time spent, speed and number of pages of the corrections to the responses by the evaluator LLMs.

Evaluator	Characters	Time Spent [min]	Speed [characters/min]	Pages
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
Mean	184406	98	1865	**Total**
R.S.D. [%]	42%	19%	36%	1484

2.2.3. Evaluation of the Responses

In general, the evaluation procedure adopted by the LLMs was similar for all of them: for each corrected response, a table was generated with the scores corresponding to the requested evaluation items, justifications for the scores given, a general summary of the evaluation, a list of the strengths and weaknesses of the LLM candidate who provided that answer, as well as an analysis of whether the answer would be

compatible with a candidate for industrial supervisor. Some candidate LLMs adopted an informal style in their responses, a fact that was criticized by the evaluating LLMs. Table 6 shows the scores received by the LLM candidates according to the items adopted by the evaluating LLMs, as well as their overall evaluation. It can be observed that the variation between the final scores was not very high, ranging from 4.4 to 4.8 for a maximum of 5.0. Claude was the LLM who performed best overall, receiving a score of 4.82, which was to be expected given the high volume of information and minimal number of hallucinations in his responses. Curiously, second place went to Gemini, the least verbose LLM in the series, who received a score of 4.45. DeepSeek took third place, with a score of 4.38, closely followed by ChatGPT, who received a score of 4.36. Finally, last place went to CoPilot, with a score of 4.26.

Table 6: Grades for the responses given for each of the LLMs to the questionary, according to the evaluation parameter and the overall mean.

Candidates → Criterion ↓	ChatGPT	CoPilot	Gemini	Claude	DeepSeek	Mean	R.S.D. [%]
Technical Correctness	4.3	4.1	4.1	4.8	4.1	4.3	7%
Scope & Relevance	4.6	4.7	4.8	5.0	4.9	4.8	4%
Metallurgical Coherence	4.5	4.2	4.3	4.8	4.2	4.4	6%
Reasoning Coherence	4.5	4.4	4.8	4.9	4.5	4.6	5%
Logical Consistency	4.6	4.5	4.7	4.8	4.3	4.6	4%
Technical Terminology	4.6	4.5	4.4	5.0	4.6	4.6	5%
Clarity & Depth	4.1	4.2	4.9	5.0	4.8	4.6	9%
Objectivity	4.1	3.8	4.2	4.2	3.7	4.0	6%
Completeness	3.6	4.0	3.9	5.0	4.5	4.2	13%
Hallucination Rate	4.9	4.3	4.5	4.7	4.2	4.5	6%
Overall Mean Score	4.36	4.26	4.45	4.82	4.38		
R.S.D. [%]	8%	6%	8%	5%	8%		

The same table shows that the worst-rated evaluation item, for all LLMs, was Objectivity, with an overall mean of 4.0, mainly in the case of DeepSeek and CoPilot. This is probably due to the tendency of all of them to exaggerate the detail in their responses. Interestingly, the second worst score (4.2) was in the Completeness item, where several LLM candidates were criticized for their evaluating LLMs for not including details that they considered important in the responses, claiming that they adopted a textbook style and did not cite practical or industrial details. Another frequent criticism was the lack of a quantitative approach in responses to the questions. A lot of subjectivity was found in this analysis, as the scores for this item showed a maximum relative standard deviation of 13%, since the scores ranged from 3.6 for ChatGPT to 5.0 for Claude, the latter being the case where the responses had maximum detail. Technical Correctness received the third worst score (4.3), closely followed by Metallurgical Coherence (4.4), with only Claude being well-rated (4.8). The Scope & Relevance assessment item was the best rated, with an overall mean of 4.8 and a minimum relative standard deviation of 4%; the other items received intermediate scores.

Table 7 shows the scores given by the evaluating LLMs to the LLM candidates. The high variation in the overall mean assigned by the evaluating LLMs – from 3.7 (Claude) to 4.9 (Gemini) – shows a significant level of subjectivity among them. In fact, Claude was the most severe evaluator because, since it generated the most complete and detailed responses to the questionnaire, it gave poor evaluations to the other LLMs, whose responses were much less comprehensive. The exception was Gemini, who, despite having the shortest responses observed in this series, was the best evaluated by Claude, receiving a score of 4.0, although not very high. Note that ChatGPT was the second LLM that made the most severe corrections in the series. The others were more lenient, especially Gemini, which assigned a score of 4.9 to all its competitors. Curiously, DeepSeek assigned the maximum score to Claude in all evaluation criteria, which, naturally, was reflected in its final score.

Table 7: Grades given by the LLM evaluators to the LLM candidates.

Candidates → Evaluators ↓	ChatGPT	CoPilot	Gemini	Claude	DeepSeek	Mean	S.D. [%]
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%
Mean	4.4	4.3	4.5	4.8	4.4		
S.D. [%]	15	14	8	3	12		

As shown in Table 8, the hiring of LLM candidates Claude and Gemini was unanimous among the evaluators. ChatGPT came in third place, with Claude recommending his hiring for a twelve-month internship in the role plus a new evaluation, and DeepSeek unconditionally vetoing his hiring. These recommendations were consistent with the scores obtained by the candidates in their respective responses to the questionnaire.

Table 8: Verdicts of the LLM candidates.

Evaluators → Candidates ↓	ChatGPT	CoPilot	Gemini	Claude	DeepSeek	Verdict
Chat GPT	-	Yes	Yes	Probation	No	No
CoPilot	As Trainee	-	Yes	Oral Exam	No	No
Gemini	Yes	Yes	-	Yes	Yes	Yes
Claude	Yes	Yes	Yes	-	Yes	Yes
Deep Seek	As Trainee	No	Yes	Oral Exam	-	No

2.2.4. Quality of the Answers and Corrections

The metallurgical analyses performed by the LLMs in their responses and corrections were very numerous and detailed; describing them would require writing a lengthy article dedicated exclusively to this topic. However, it is possible to make some comments on the most significant and recurring cases.

In question 6, ChatGPT confused Recrystallization Limit Temperature (RLT), Recrystallization Stop Temperature (RST), and Temperature of No-Recrystallization (T_{nr}). CoPilot detected this problem. But the fact is that this type of confusion is also common in established technical literature.

In the same question, ChatGPT, CoPilot, Gemini, and Deep Seek initially stated that the theoretical minimum strain for austenite recrystallization would be in the range of 0.10 to 0.15 – in reality, these are values that are too low. However, they then contradicted themselves, stating that in industrial rolling this range would be between 0.15 and 0.20, which is more accurate. Curiously, despite having made the same mistake in their answers, this inconsistency was pointed out by CoPilot and DeepSeek in their corrections! Additionally, it was also found that these LLMs failed to adequately distinguish the differences between dynamic and static recrystallization.

Some LLMs, such as CoPilot and DeepSeek, included links to technical articles in some of their responses and corrections; the content of these articles was likely used to generate those texts. Claude, however, only cited authors involved in the topics discussed or their respective books. It criticized the use of links in responses, in some cases considering it inappropriate for inclusion in formal technical evaluations. In the specific case of DeepSeek, Claude argued that the highly specific information present in those articles did not allow for valid answers in general terms, as requested in the questionnaire.

But even Claude, who was so highly rated, made mistakes. In only two cases it and DeepSeek committed elementary errors when calculating the sum of the scores for answers to questions. Claude also made more serious errors, contradicting himself on a metallurgical question – more specifically, the effect of the strain rate on the minimum recrystallization temperature. In another case, it criticized the use of the so-called "shape factor" parameter in one of the questionnaire answers but, a few answers later, praised its use. And, in fact, this factor is fundamental for determining pass schedules that guarantee the internal soundness of heavy plates.

Another criticism, mainly made by Claude in relation to the other LLMs, is the textbook tone that dominated the questionnaire responses. This seems to indicate that the learning sources for LLMs should not include practical, industrial-level publications on controlled lamination.

Interestingly, ChatGPT spontaneously stated, at the end of its evaluations, that the questionnaire responses prepared by CoPilot and Claude were highly likely to have been generated by artificial intelligence, due to their didactic style, numerous items, and large size. CoPilot made the same comment about DeepSeek's responses. According to the evaluating LLMs, this finding, if confirmed, would constitute a demerit for the "candidates".

Finally, it is necessary to observe that the approach adopted here, where static responses to the questionnaire were obtained and automatically corrected by the LLMs, is not the best way to obtain correct answers. Much better and less uncertain results are obtained using the ‘interactive prompting’ technique, i.e., through the development of a progressive and critical conversation between the user and the LLM, which generally promotes the emergence of very useful insights, even if they have to be subsequently confirmed in the established literature [7]. Another option is the so-called prompt “chain of thought”, where the LLM must present its reasoning step by step before elaborating its conclusion. In this way, the LLM can act in roles where it performs best, i.e., in the organization of reasoning and compilation of knowledge.

3 CONCLUSION

This study has demonstrated, somewhat unexpectedly, that public LLM-based chatbots have shown relatively satisfactory mastery of the subject of controlled rolling of microalloyed structural steels, even in their free version. For those who, a few decades ago, had great difficulty obtaining sources of knowledge on this subject in books and journals with restricted access, it is very gratifying to see that, nowadays, students, as well as technicians and engineers at the beginning of their careers, can obtain reasonably good quality information quickly, free of charge, and anywhere with an internet connection.

The overall assessment of the LLMs' knowledge in the area studied here did not vary much, with scores ranging from 4.3 (out of a maximum of 5.0) for CoPilot to 4.8 for Claude. The remaining platforms, ChatGPT, Gemini, and DeepSeek, fell into the intermediate group. Gemini, with a score of 4.5, came in second place and is a good option for those who want to obtain reasonably high-quality knowledge in a concise way, without going into the excessive detail that characterized Claude's answers.

But it should be remembered, once again, that the results generated by LLMs are not deterministic, i.e., they fluctuate statistically. They are remarkably similar to Schrödinger's cat, which is simultaneously alive and dead — it only assumes a definite state at the moment of observation. There is no guarantee that the same result will be obtained again.

So, all the information provided by LLMs must be critically analyzed, as should always be done when seeking knowledge in specialized literature. Many of the flaws observed in this study originated from sources developed by humans that served as the basis for the LLMs' learning. On the other hand, it cannot be forgotten that LLMs do not reason; they only generate very plausible texts. Great care must be taken if the information obtained from LLMs is used in an academic or professional environment. It is recommended here not to limit oneself to the first answer obtained when consulting an LLM. Artificial intelligence should be an intellectual interlocutor rather than an authoritative source – i.e., human must be in the loop elaborating the interactive prompts. Sometimes the journey can be more important than the destination.

REFERENCES

1 KLESEL, M.; WITTMANN, H.F. Retrieval-Augmented Generation (RAG). Business & Information Systems Engineering, vol. 67, n.4, p. 551–561, 2025.

2 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.

3 CHEN, B. et al. Unleashing the potential of prompt engineering for large language models. Patterns, vol. 6, issue 6, 44 p., June 13, 2025.

4 BHAVSAR, P. Mastering RAG. Burlingame, E.U.A.: Galileo AI Inc., 2026, 194 p.

5 HOSKING, T.; BLUNSOM, P.; BARTOLO, M. Human Feedback is Not Gold Standard. arXiv:2309.16349 [cs.CL], 2024.

6 BHAVSAR, P. Mastering LLM-as-a-Judge. Burlingame, E.U.A.: Galileo AI Inc., 2025, 70 p.

7 MIDOLO, A. et al. Guidelines to Prompt Large Language Models for Code Generation: An Empirical Characterization. arXiv:2601.13118 [cs.SE], 2026.