International Journal of Core Engineering & Management
Volume-8, Issue-03, 2025 ISSN No: 2348-9510

IMPROVING TRUST ON GEN AI - A BIDIRECTIONAL APPROACH

Anand Athavale
andyathavale@gmail.com
Independent Researcher,

Abstract

Gen AI has improved quite a lot over the years. However, one who have critical thinking mind, warn everyone to use Gen AI in "Trust but Verify" approach. The primary reasons being hallucinations and bias which is not necessarily Gen AI model shortcoming, but more to do with the input data and quality. Latest developments such as advanced reasoning can help to some extent but those too are still at the mercy of knowledge and information fed into the model. The concept of confidence score does not guarantee accurate prediction. The answer to trust may be found by combining two focus areas. One, the old AI programming language constructs like Prolog by using those as cross-checking mechanisms instead of problem-solving methods, and mimicking the correct human behavior of asking for more information when in doubt. Second, looking at human behavior and roots for building trust.

Keywords— Gen AI, Hallucinations, Confidence Score, Increase adoption, Trust Building

I. INTRODUCTION

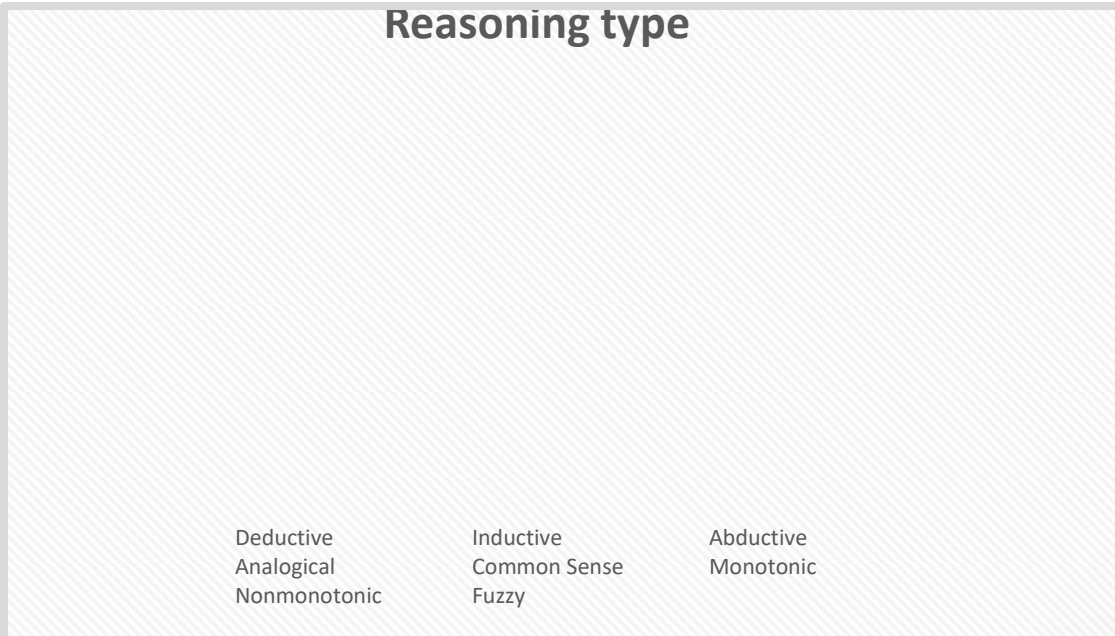
Generative AI is at a point where improvements have shifted direction to improve adoption instead of improving speeds. The features like advances reasoning are mainly geared towards improving the response or prediction quality instead of speed. Built in concepts like confidence scores and levers like temperature to control randomness have a clear sealing when it comes to accuracy and usefulness, which eventually would result in trusting a Gen AI response. Only handful of information is not susceptible to bias or hallucinations. If you ask "In which direction does the sun rise," the answer is always East. Unfortunately, this confidence and accuracy fades for more specific questions which are not based on a single piece of information, which the model would have been trained on. Lastly, all the efforts being put are on making Gen AI think like humans. But zero effort is being put on studying human behavior to find avenues of increasing trust.

II. AI REASONING IMPACT ON TRUST INCREASE

AI reasoning is somewhat latest addition in Gen AI. It was primarily done to improve AI reliability and trustworthiness [1]. There are the various types of reasoning [2]. Each adds value but does not improve the trust part to the level that humans would be comfortable with. To be clear, this is not to say that AI reasoning is not required, or is not useful. It is most definitely

required. But there are gaps which need to be highlighted, to then look for complimentary ways to fill those gaps.

1. **Deductive –** Oversimplified, these are based on rules like if A means B and B means C, then A means C. It assumes A is a fact.
2. **Inductive –** This is more like behavior patterns-based reasoning. If A has shopped for groceries on Saturday morning for the past 3 years, it is likely that A will go to shopping next Saturday. There are always anomalies which could break a pattern.
3. **Abductive –** Oversimplified, this is closer to "if it looks like a duck, swims like a duck, and quacks like a duck, then it's probably a duck." This does not give full confidence in the conclusion.



Various Reasoning Types

4. **Analogical –** Oversimplified, this is closer to "a car with manual gears is like a bus, so anyone who can drive a car with manual gears can drive a bus" type conclusion. Clearly, this does not instill confidence.
5. **Common Sense –** Using a humor-based example, this reasoning is, "if you cut the very branch from the outer end on which you are sitting, you will fall" type reasoning. This is useful and builds trust, but opportunities to make use of these are rare, especially in complex problem solving. Additionally, not all common-sense scenarios would be clearly documented for Gen AI models to learn from it.
6. **Monotonic –** These are basically irrefutable facts. To give a mathematical example, "Square has four sides" will not change even if one adds more information such as "size of each side for a square is 4 inches". Non-mathematics example could be "Bulb with filaments heats

when turned on" will not change even if one says "brightness measurement of the bulb is only 800 lumens". Among all reasoning, this type of reasoning is absolute guarantee of trust, but same as common sense reasoning, the use of Monotonic reasoning may not be possible for all problems or prompts.

7. **Non monotonic** – These are rare exceptions to various other type of reasoning. In school, we learnt about such things. Example is "All mammals give birth directly to young living children." The well-known exception to that is "platypus." This is the type of reasoning which has potential of generating "infinite" exception scenarios, which can reduce confidence score of responses [3] and in turn the trust on Gen AI.
8. **Fuzzy** – This is where "generic" nature of a statement to be true is considered acceptable instead of "specifics" of it. The example of this may be a statement "Careful, the pot is hot." In general, the statement would be true and it is made to imply that "There is high percentage of chance that if you touch the pot, you will burn your fingers." But it does not state what is the current temperature, neither gives specifics for contact burning temperatures [4].

III. OTHER CONTROLS FOR TRUST AND RELIABILITY

Besides advances reasoning, there are transparency and control measures which could improve trust. But each has flaws of their own. Again, it does not mean these are not useful. It is just that these require knowledge and comfort with Gen AI workings to increase trust.

1. Confidence Score

A confidence score, in essence, is a numerical representation of how sure an AI model is about its prediction. It is typically a value between 0 and 1 (or 0% and 100%), where close to 1 Indicates high confidence that prediction is correct and 0 indicates low confidence indicating uncertainty about its prediction [5]. Score around 0.5 or 50% is indicative of guessing or hallucinations. These in turn depend on calibration to make confidence score more reliable. In other words, even if reliability is indicated, reliability of the reliability score itself is not absolute.

2. Temperature

The temperature control in Gen AI decides whether a response will be more predictable, conservative text (response), or more varied and sometimes more creative or unexpected text (response) [6].

For example, if there are areas of the problem left out in a solution description, a high temperature control prompt asking about verifying capabilities of a solution, might result in inaccurate representation in the response for the missing part and low temperature may result in admittance of inability of verification.

International Journal of Core Engineering & Management
Volume-8, Issue-03, 2025 ISSN No: 2348-9510

IV. THE "HUMAN" ANGLE OF TRUST

There a couple of things to note about how humans approach trust from a psychological point of view.

1. Default Trust behaviour

Humans have changed over the time related to their default trust behaviour. It also varies in personal and professional set ups. The default behaviour in professional workspaces is to not trust a new comer immediately. It is either a wait and watch approach, or, many times plagued with preconceived "hear-and-say" about the person. Once you look consider this, human behaviour toward Gen AI is would not be that different.

2. Trust improvement through interaction

Time and again, companies find that the best way, and if I may say the only one way, to improve trust is through interaction and collaboration. Keeping the hierarchies and authority in mind, the trust still requires a good amount of interaction.

Overall, these two aspects highlight the focus distribution needed in the pursuit of building trust in Gen AI. Focusing only on making Gen AI be more like human is a single-minded liner approach. The "objective" of Gen AI should shift from generating a correct answer to "being a contributor, collaborator and team player" so that the desired outcome is achieved "together" with the human.

V. POSSIBLE METHODS TO MAKE GEN AI INTERACTIVE

1. Clarifying questions

Gen AI already has this capability. However, the questions are more geared and asked by Gen AI with focus on eventually giving the answer. A shift from this objective is needed so that, the questions are generated to indicate inadequate knowledge. Instead of asking more questions around objectives like "Do you mean X or Y", the questions should be also asked by Gen AI such as "I am not sure I have enough information from trusted sources about this sub-objective. Would you be able to answer this for me?" This is the approach suggested here about Gen AI built to ask more refinement of the question [7].

2. Asking for decisions in sub-step stage instead of only as follow-ups

Typically, Gen AI asks about decision or opinion only after giving the entire response. Instead, Gen AI should make use of old programming language paradigms of facts and rules [6]. Here, the shift is to use these constructs for cross-checking and collaborations instead of objective achievement. The concepts of confidence scores and calibration could be fitted into the facts and rules constructs to decide on "pausing the work on the sub-objective" and asking for continuance with some assumption, or, a choice between multiple options. To illustrate, let's say you have 10,000 daily data points as "Date and temperature" and then asked Gen AI to generate a new column as "Daily churn" by doing a difference between the date and the previous date temperature. As it happens, say May 18th 1999 and August 12th 2008 were

absent. Gen AI should not simply ignore it and calculate value of May 19th as May 19th temperature minus May 17th temperature. The trust building approach would have Gen AI halt the calculation, prompt a question to the human asking something like "I do not have May 18th data. Was it accidentally removed? How would like me to proceed? Should I just skip, or calculate a predicted value for May 18th 1999 temperature and then add that record?"

To be clear, this example is for a deterministic objective to illustrate "pause and ask question or guidance" behavior. Gen AI may be doing something like this already with some control settings. The trust improvement is needed on more non-mathematical, research type query responses.

3. Authority delegation

In certain situations, the best way to earn trust in the response is to follow the aviation CRM (Crew Resource Management) model. Depending on situational awareness, the co-pilot (Gen AI) should be able to indicate error in the captain's (human) observation. At the same time, the captain (human) should be aware of what is known and what is unknown and communicate the same to the co-pilot (Gen AI) to transfer the decision-making authority in the "pause and collaborate" interactions when generating a response.

In other words, while the efforts are to make "Gen AI" behave like humans, humans also may need to treat Gen AI like human co-workers, on certain occasions.

VI. CONCLUSION

Trust improvement is needed to not only increase the reliability of the responses but adoption of Gen AI for non-deterministic problems or query responses. Existing techniques are moving the needle but it has not reached the reliability level which would make a responsible inquisitive human comfortable with the Gen AI response. There are more efforts to make Gen AI like humans with a belief that doing so would get Gen AI closer to become reliable. However, certain amount of focus is needed to understand human trust building with other humans to balance the need to answer everything no matter what and to transfer sub-step control back and forth between Gen AI and humans depending on the awareness and confidence of both for each sub step of generating a reliable response.

REFERENCES

1. Danilo Poccia, Automated reasoning and generative AI: Harness creativity with formal verifications (Jan, 2025), https://dev.to/aws/automated-reasoning-and-generative-ai-harness-creativity-with-formal-verifications-o6 , (April, 2025)
2. Bhushan Jadhav, AI Reasoning Explained, (January, 2024), https://aisera.com/blog/ai-reasoning/ , (April, 2025)
3. Sampurna Mandal, Ankush Ghosh, Single shot detection for detecting real-time flying objects for unmanned aerial vehicle, Artificial Intelligence for Future Generation Robotics ,

(2021), https://www.sciencedirect.com/topics/computer-science/confidence-score , (April, 2025)
4. ANTISCLAD, Burn Exposure Chart, (2016), https://antiscald.com/index.php?route=information/information&information_id=15 , (March, 2025)
5. Alphanome.ai, Understanding Confidence Scoring in AI, (January, 2025), https://www.alphanome.ai/post/understanding-confidence-scoring-in-ai , (April, 2025)
6. Emily Lewis, MS, CPDHTS, CCRP, Setting the AI Thermostat: Understanding Temperature to Balance Creativity and Coherence, (January, 2024), https://www.linkedin.com/pulse/setting-ai-thermostat-understanding-temperature-emily-rh8qc , (April, 2025)
7. Gianni Giacomelli, Researcher | Consulting Advisor | Keynote | Chief Innovation / Learning Officer, GenAI must ask questions, not just give answers (2016), https://www.linkedin.com/pulse/genai-must-ask-questions-just-give-answers-gianni-giacomelli-yrhaf , (April, 2025)