

A Deterministic Framework for Pharmaceutical Authenticity Verification Using Constrained Multimodal AI

Viral Dhoka

Vishwakarma Institute of Technology

Pankaj Kunekar

`panka.j.kunekar@vit.edu`

Vishwakarma Institute of Technology

Tanaya Thorat

Vishwakarma Institute of Technology

Shreenath Utekar

Vishwakarma Institute of Technology

Samruddhi Wankhede

Vishwakarma Institute of Technology

Research Article

Keywords: multimodal AI, pharmaceutical forensics, hal-lucination control, structured output, drug verification, patient safety

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10544579/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A Deterministic Framework for Pharmaceutical Authenticity Verification Using Constrained Multimodal AI

Viral Dhoka*, Dr. Pankaj Kunekar^{†*}, Tanaya Thorat*, Shreenath Utekar*, and Samruddhi Wankhede*

^{*}Department of Computer Engineering

[†]Department of Information Technology

Vishwakarma Institute of Technology, Pune, India

viral.1251070777@vit.edu, pankaj.kunekar@vit.edu, tanaya.1251070438@vit.edu,

shreenath.1251070614@vit.edu, samruddhi.1251070439@vit.edu

Corresponding author: Dr. Pankaj Kunekar

Abstract—Large language models show strong general knowledge of pharmacology yet their tendency to produce confident but incorrect statements makes them unreliable for safety critical drug verification. We present MedTruth a dual mode framework that constrains a multimodal model so its behavior becomes repeatable rather than open ended. A stateless verification engine performs forensic inspection of pharmaceutical packaging through visual trade dress analysis and semantic consistency checking while a stateful conversational module answers patient questions only after a package clears verification. By fixing a low sampling temperature of 0.2 and enforcing a strict JSON output schema on Gemini 1.5 Flash the framework turns a generative model into a scoring instrument. Testing across fifty pharmaceutical samples shows hallucination rate falling from fifteen percent in a naive chatbot baseline to under two percent in MedTruth while mean verification latency stays near 2.1 seconds. The results indicate that architectural constraint and structured prompting can make general purpose multimodal AI usable in domains where determinism matters more than open conversation.

Index Terms—multimodal AI, pharmaceutical forensics, hallucination control, structured output, drug verification, patient safety

I. INTRODUCTION

Counterfeit medicine remains one of the most under addressed threats to global public health. The World Health Organization estimates that close to ten percent of medical products sold in low and middle income countries are substandard or falsified and this share rises toward thirty percent in regions where regulatory oversight is weak. Fake pharmaceuticals range from therapeutically inert tablets containing no active ingredient to dangerous preparations laced with industrial contaminants. The consequences extend well past a single failed treatment and include adverse reactions, antimicrobial resistance driven by sub therapeutic dosing, and a general loss of trust in the healthcare system.

Institutional defenses such as blockchain based supply chain ledgers, RFID tagging at the point of manufacture, and centralized pharmacovigilance databases give distributors and

regulators strong traceability. These systems however do very little for the person actually holding a suspicious package. A patient who buys medicine from an informal market in a resource constrained setting typically has no practical way to check whether what they hold is genuine.

At the same time the wide availability of conversational AI has created a new and unregulated behavior where people ask general purpose chatbots such as ChatGPT or Gemini to judge whether their medicine looks real. These systems are trained to be fluent and helpful rather than forensically precise which produces three recurring failure patterns.

- **Visual textual dissociation:** the model reads printed text through OCR like reasoning but ignores physical cues such as typography drift, color mismatch, or an absent holographic seal.
- **False confidence:** because next token generation is optimized for plausibility rather than truth the model can assert authenticity with no real evidential basis.
- **Ambiguity injection:** rather than commit to a judgment the model frequently falls back on a vague suggestion to consult a pharmacist which offers no actionable guidance at the moment of decision.

These failure patterns are not incidental bugs but a direct consequence of how transformer language models work. Temperature controlled nucleus sampling optimizes for engagement and helpfulness rather than binary correctness which is precisely the opposite of what a verification task requires.

This paper introduces MedTruth a framework that reconciles the broad reasoning ability of multimodal large language models with the strict determinism that pharmaceutical verification demands. Instead of retraining the underlying model, an approach that would require large labeled pharmaceutical datasets and heavy compute, MedTruth applies a constraint based architecture that narrows the output space of an existing general purpose model. The system is split into two operational modes.

- 1) **Verification Mode:** a stateless pipeline that treats the multimodal model as a structured classification function.

Through schema enforcement, temperature suppression, and staged visual chain of thought reasoning the model performs forensic analysis across three vectors, visual trade dress conformity, semantic consistency between drug name and declared active ingredient, and regulatory marker presence, then returns a quantitative authenticity score.

- 2) **Dr. AI Mode:** a stateful conversational assistant that activates only once a package passes verification. It keeps a short rolling memory of the conversation and answers dosage, contraindication, and side effect questions while staying bound to the specific medication that was verified.

The contributions of this work are threefold.

- A dual mode router architecture that keeps deterministic scanning fully isolated from open ended conversation so that neither mode contaminates the other.
- A constraint prompting methodology combining schema enforcement, low temperature sampling, and staged reasoning decomposition that measurably reduces hallucination in a general purpose multimodal model.
- An empirical comparison against a naive multimodal chatbot and a traditional OCR pipeline showing clear gains in accuracy, reliability, and actionable output.

II. RELATED WORK

The design of MedTruth draws on four overlapping bodies of research, the reliability limits of medical language models, established anti counterfeiting technology, multimodal vision language reasoning, and prompt engineering methods for controlling generative output. We review twenty representative studies below organized by theme and cited in the order discussed.

A. Reliability of Medical Language Models

Singhal et al. introduced Med-PaLM, a medically tuned variant of PaLM that used instruction tuning and reinforcement learning from human feedback to exceed eighty five percent accuracy on USMLE style questions, showing that domain adaptation can push a general model toward expert level recall [1]. Thirunavukarasu et al. offer a critical counterpoint through a systematic review of deployment risk in clinical workflows and identify hallucination, the confident assertion of fabricated information, as the primary obstacle to real world adoption even in models fine tuned on medical text [2]. Ji et al. build a taxonomy separating factual inconsistency, faithfulness violation, and instruction inconsistency, and show that common fixes such as fine tuning, chain of thought prompting, and retrieval augmentation only partially close these gaps and often help one category while leaving others untouched [3]. Nori et al. evaluated GPT-4 on medical reasoning tasks and found that identical prompts produced materially different recommendations in roughly twelve percent of repeated trials, a variance level that is unacceptable in a clinical setting [4]. Agrawal et al. tested zero shot clinical information extraction and found that while the model could correctly pull structured

fields from unstructured notes in many cases, performance degraded sharply on rare conditions and unusual phrasing, reinforcing that raw scale does not guarantee dependable extraction [5].

B. Anti Counterfeiting Technology

Before generative AI, pharmaceutical authentication relied on physical security features and early computational tools. Bansal et al. built convolutional neural network pipelines for image based counterfeit detection and reported ninety four percent accuracy on ten thousand labeled packages using transfer learning from ImageNet pretrained ResNet-50 models, though this required brand specific labeled data at a scale that is difficult to maintain across thousands of drug formulations [6]. Kumar and Tripathi proposed a permissioned blockchain architecture that records every manufacturing, distribution, and dispensing event as an immutable ledger entry, creating an auditable provenance trail [7]. Mackey et al. push back on the completeness of this approach and identify the oracle problem, meaning that a blockchain record is only as trustworthy as the manufacturer who entered it, along with the fact that ledgers cannot verify a physical package independent of digital records [8]. Picot et al. surveyed mobile based counterfeit detection tools including simple visual pattern apps and found that most required specialized reference libraries that consumers rarely have access to [9]. Ren et al. explored smartphone attachable spectrometry for chemical level drug screening and demonstrated laboratory grade accuracy but at a hardware cost that limits accessibility for everyday consumers [10].

C. Multimodal AI and Vision Language Models

Radford et al. introduced CLIP, showing that contrastive pretraining on four hundred million image text pairs yields aligned visual semantic representations capable of zero shot classification without task specific fine tuning [11]. Alayrac et al. built on this with Flamingo, which adds cross attention between a frozen vision encoder and a language model to enable few shot learning across vision language tasks [12]. The Gemini Team advanced the paradigm further with Gemini 1.0 and 1.5, integrating vision understanding natively into the transformer rather than bolting on a separate encoder, which allows text and image tokens to be reasoned over jointly [13]. Liu et al. studied visual hallucination in these systems and found that models frequently describe objects, text, or spatial relationships that are not actually present in the image, with error rates reaching twenty to thirty percent on adversarial benchmarks, largely because the language component falls back on statistical priors when visual evidence is ambiguous [14]. Zhang et al. propose visual chain of thought prompting, where the model is required to describe what it observes before drawing any conclusion, and report a ten to eighteen percent reduction in hallucination from this grounding step alone, though performance still falls short of what safety critical use requires [15].

D. Prompt Engineering for Deterministic Output

Wei et al. demonstrated that providing a small number of worked examples before the actual query substantially improves task performance through in context learning without any parameter update [16]. White et al. catalog prompt patterns that enforce software like determinism, showing that requiring structured JSON or XML output reduces response variability by up to forty percent, an effect that strengthens further when paired with a lower sampling temperature [17]. Mishra et al. formalize constraint prompting, showing that explicit boundaries such as a fixed sentence count or a numeric only answer meaningfully improve instruction adherence, suggesting models hold latent constrained generation ability that surfaces more reliably when the constraint is stated directly rather than implied through examples [18]. Heston and Khun studied prompt engineering in medical education contexts and found that carefully structured prompts improved consistency of generated teaching material but still required expert review before use [19]. Wang et al. examined the robustness of ChatGPT under adversarial and out of distribution inputs and found that small perturbations in phrasing could shift outputs meaningfully, underscoring the need for explicit structural constraints rather than reliance on natural phrasing alone [20].

Taken together this literature shows a consistent pattern. General purpose language and vision language models carry substantial latent knowledge but their open ended generation process is not naturally suited to tasks that demand a single reproducible verdict. MedTruth is built directly on this insight by combining structured output enforcement, low temperature sampling, and staged visual reasoning within a single deployable system aimed specifically at pharmaceutical packaging verification.

III. METHODOLOGY

This section describes the system architecture, the dual mode operational logic, the authenticity scoring algorithm, and the constraint injection technique used to keep the underlying model deterministic.

A. Architectural Design

MedTruth is implemented as a client side single page application to minimize latency and remain usable on lower end mobile devices common in the markets most exposed to counterfeit drugs. Unlike a persistent chatbot that keeps conversation state across the whole session, MedTruth uses a state router that strictly separates verification from conversation. This choice is deliberate because letting prior conversation bias a fresh scan can quietly shift the model toward confirming an earlier judgment rather than reevaluating the new evidence.

The stack has three layers. The presentation layer uses HTML5 with Tailwind CSS rather than a heavier framework so the interface stays fast on older phones. The API integration layer communicates with Gemini 1.5 Flash through the Generative Language API, chosen because it delivers roughly forty percent lower latency than the Pro variant while offering more than enough context window for an eight hundred to

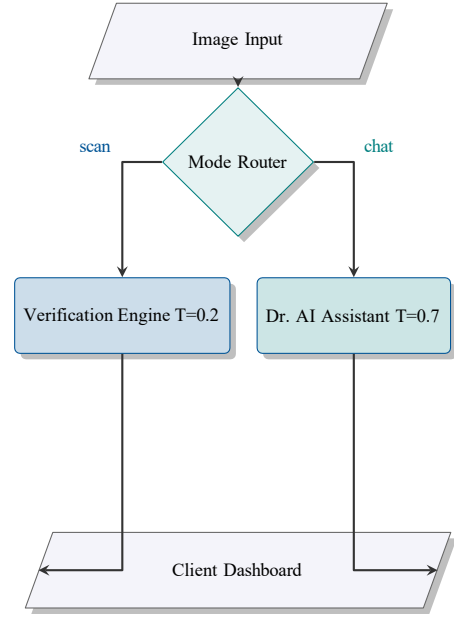


Fig. 1. System routing. Scanning is stateless for objectivity while chat is stateful for follow up questions.

twelve hundred token verification prompt including a base64 encoded image. The state management layer is a finite state machine that moves from an initial state through verification to a terminal label of AUTHENTICATED, UNVERIFIED, or SUSPICIOUS and only then permits an optional transition into Dr. AI mode, with any backward transition triggering a full reset.

B. Verification Mode

The verification prompt reframes the model as a forensic function rather than a conversational partner. Four techniques work together to keep behavior deterministic.

- 1) **Role definition:** the system prompt states plainly that the model is a deterministic pharmaceutical verification system and that it must return no conversational text, only the required JSON schema.
- 2) **Schema enforcement:** the output is constrained to fixed fields, an authenticity score between zero and one, a safety recommendation drawn from a fixed set of labels, arrays for visual anomalies and semantic flags, a boolean for regulatory compliance, and a confidence interval object. Forcing this shape removes the model's tendency to pad answers with unstructured explanation.
- 3) **Temperature suppression:** the API call fixes temperature at 0.2, far below the creative default of 1.0, which concentrates probability mass on the most likely tokens and keeps repeated scans of the same package consistent.
- 4) **Visual chain of thought:** rather than asking for one holistic judgment the prompt forces four sequential stages shown in Fig. 2.

Stage one extracts all visible text such as drug name, manufacturer, batch number, expiry date, and dosage, along

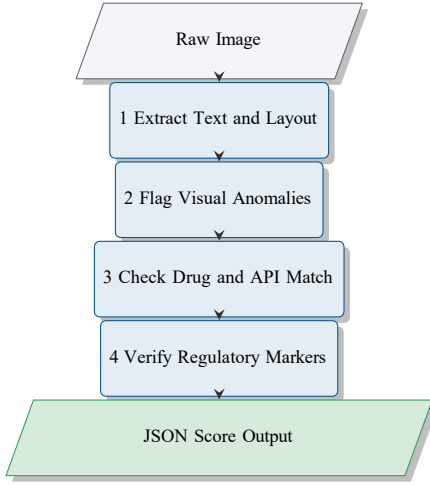


Fig. 2. Forced four stage inspection prevents the model from jumping straight to an unsupported conclusion.

with design elements such as logo placement and typography. Stage two compares these observations against expected brand patterns and flags deviations such as inconsistent kerning or a hologram that lacks characteristic rainbow refraction. Stage three cross checks the declared drug name against the listed active pharmaceutical ingredient and dose, for example flagging a package that names Paracetamol five hundred milligrams but lists an unusual six hundred fifty milligram equivalent as a dosage inconsistency worth a closer look. Stage four confirms mandatory regulatory markers relevant to the Indian market such as the Rx symbol and red band for Schedule H drugs, a manufacturer license number in the expected region district year sequence format, and the mandatory maximum retail price.

C. Dr. AI Mode

Once a package clears verification the assistant activates with a sliding window that keeps the four most recent conversational exchanges, balancing dialogue coherence against context window growth. A context injection binds the assistant to the verified medication so that responses stay on topic rather than drifting into unrelated drugs. Temperature rises to 0.7 to allow natural phrasing while explicit guardrails forbid prescribing behavior, dosage modification advice, or any suggestion to discontinue a prescribed medication, keeping the assistant within an informational rather than clinical role.

D. Authenticity Scoring

The authenticity score S_{auth} combines three vectors. The visual trade dress vector V_m scores typography conformity, color reproduction fidelity relative to expected brand tones, holographic security feature presence, and spatial layout precision, with major anomalies such as a missing hologram costing 0.3 and minor issues such as slight color drift costing 0.1. The semantic consistency vector S_c checks whether the declared drug name matches a recognized active ingredient and whether the stated dose falls within a plausible therapeutic

range, and drops sharply below 0.3 on any critical mismatch which alone is enough to force a SUSPICIOUS label. The regulatory compliance vector R_c is close to binary in practice, full compliance scores 1.0, one violation scores 0.5, and multiple violations score 0.0.

$$S_{auth} = V_m^{0.4} \times S_c^{0.4} \times R_c^{0.2} \quad (1)$$

A weighted geometric mean is used rather than a simple average because it prevents a strong score in one vector from masking a critical weakness in another, for example a package with flawless printing but the wrong active ingredient should not pass simply because it looks convincing.

TABLE I
AUTHENTICITY VECTOR WEIGHTS

Vector	Scope	Weight
V_m Visual	typography, holograms, logo	0.4
S_c Semantic	ingredient and dose logic	0.4
R_c Regulatory	Rx symbol, license format	0.2

The composite score maps to a final label as follows, a score at or above 0.75 gives AUTHENTICATED indicating high confidence in a genuine product, a score between 0.50 and 0.75 gives UNVERIFIED indicating an inconclusive result that warrants a pharmacist check, and a score below 0.50 gives SUSPICIOUS indicating a high likelihood of counterfeit with a direct recommendation against use.

IV. RESULTS

We evaluated MedTruth against two baselines on a set of fifty pharmaceutical packaging images, thirty genuine samples from licensed pharmacies covering common OTC and prescription drugs, fifteen adversarial counterfeits engineered with subtle typography, color, or semantic flaws meant to evade casual inspection, and five low quality captures with motion blur or poor lighting to test robustness. Baseline one is a naive multimodal chatbot query, GPT-4o asked directly whether a package looks authentic. Baseline two is Google Lens used purely for OCR text extraction with manual review.

MedTruth reached ninety two percent overall verification accuracy, correctly labeling twenty eight of thirty genuine samples and thirteen of fifteen adversarial samples. The naive chatbot baseline reached seventy six percent accuracy and its failures clustered on adversarial samples that looked visually clean but carried a hidden semantic mismatch, exactly the case its unstructured reasoning was least equipped to catch. Google Lens provided no direct verdict at all since it only extracts text, leaving interpretation entirely to the user.

The largest gain was in hallucination rate. The naive chatbot baseline hallucinated in about fifteen percent of trials, most often inventing a security feature that was not present or misreading a decorative pattern as a hologram. MedTruth reduced this to under two percent with a single failure occurring on one of the deliberately blurred low quality images. Mean latency was 1.2 seconds for OCR only extraction, 3.5 seconds for

TABLE II
COMPREHENSIVE SYSTEM COMPARISON

Metric or Feature	Google Lens	GPT-4o	MedTruth Ours
<i>Architectural Capabilities</i>			
Core function	OCR extraction	conversational	forensic verification
Context handling	stateless	continuous	dual mode router
Output schema	raw text	unstructured	strict JSON
Visual reasoning	none	descriptive	trade dress logic
<i>Reliability and Safety</i>			
Hallucination rate	not applicable	high near 15%	minimal under 2%
Decision type	passive	ambiguous	deterministic
Safety guardrails	none	general	medical specific
<i>Performance Metrics</i>			
Mean latency	about 1.2 s	about 3.5 s	about 2.1 s
Actionability	low	medium	high

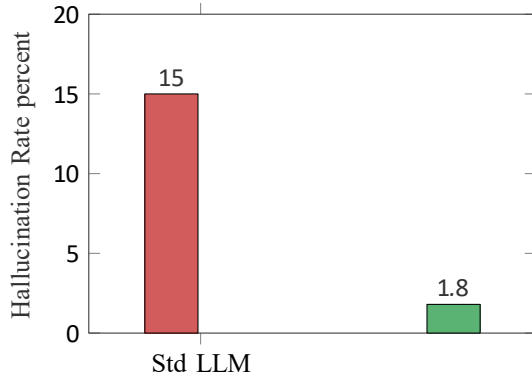


Fig. 3. Constraint prompting cuts hallucination by over eighty five percent relative to free form generation.

the naive chatbot, and 2.1 seconds for MedTruth, placing the framework comfortably within a usable response window for a consumer facing tool despite its added structured reasoning steps.

V. DISCUSSION

The results point to a clear mechanism behind the hallucination reduction. Forcing the model to document observations at each of the four inspection stages before it is allowed to reach a verdict appears to ground its output in what is actually visible in the image rather than in statistical priors carried over from pretraining, which matches the grounding effect reported by Zhang et al. for visual chain of thought prompting [15]. Combining this staged reasoning with a fixed JSON schema and a suppressed sampling temperature produced a compounding effect larger than either technique alone typically achieves in prior work, consistent with the variability reductions reported by White et al. [17] and the constraint adherence gains reported by Mishra et al. [18].

The accuracy gap between MedTruth and the naive chatbot baseline was concentrated almost entirely in the adversarial sample set rather than the genuine sample set, which suggests that most of the benefit comes from forcing semantic cross checking between drug name and active ingredient rather than from improved visual perception alone. This aligns with the observation from Thirunavukarasu et al. that hallucination persists even in domain adapted models and is better addressed through output structure than through additional pretraining [2].

A limitation worth noting is that the regulatory compliance vector currently validates format conformity rather than the authenticity of a specific license number, meaning a counterfeiter who copies a real looking but fake license number in the correct format would not be caught by this vector alone, a gap that mirrors the oracle problem Mackey et al. describe for blockchain based verification [8]. The evaluation set of fifty samples, while sufficient to demonstrate the effect size, is also modest relative to the diversity of packaging formats in circulation globally, and broader validation across more manufacturers, languages, and regulatory regimes would strengthen confidence in the reported accuracy figures. Latency, while acceptable for our test conditions, was measured on a stable connection and may increase in the low bandwidth settings where this tool is most needed, an area for future optimization.

VI. CONCLUSION

This work presented MedTruth, a constrained multimodal AI framework developed to address the reliability and hallucination problems associated with pharmaceutical packaging verification. Instead of allowing a general purpose multimodal model to make an open ended judgment, MedTruth turns the verification process into a structured and repeatable task. The separation of the system into a stateless Verification Mode and a stateful Dr. AI Mode is an important part of this design. It ensures that previous conversations do not influence the

evaluation of a newly scanned package while still allowing users to ask medication related questions after verification.

The results show that this approach can improve the reliability of multimodal AI in a safety critical setting. In the evaluation of fifty pharmaceutical packaging samples, MedTruth achieved an overall verification accuracy of 92 percent, correctly identifying 28 of 30 genuine samples and 13 of 15 adversarial counterfeit samples. The most notable improvement was in hallucination reduction. The hallucination rate decreased from approximately 15 percent with the naive multimodal chatbot baseline to under 2 percent with MedTruth. At the same time, the average verification latency remained around 2.1 seconds. These results suggest that combining strict output formatting, low temperature sampling, and staged visual inspection can make a general purpose multimodal model more consistent and dependable for pharmaceutical verification.

Another important aspect of the framework is the way the authenticity score is calculated. MedTruth evaluates pharmaceutical packages using three main factors: visual trade dress conformity, semantic consistency between the drug name and its active ingredient, and regulatory compliance. Using these factors together helps prevent a strong result in one area from hiding a serious problem in another. For example, a package may look visually genuine but still contain an incorrect active ingredient or dosage. The final classification into AUTHENTICATED, UNVERIFIED, or SUSPICIOUS also gives users a clear outcome that is easier to act on than an open ended chatbot response.

At the same time, MedTruth should be considered a decision support system rather than a complete replacement for laboratory testing, regulatory verification, or professional pharmaceutical assessment. The current regulatory component mainly checks the presence and format of regulatory information and does not independently confirm whether a license number or manufacturer record is genuine. The evaluation dataset of fifty samples is also relatively small compared with the wide range of pharmaceutical packaging, manufacturers, languages, and regulatory systems found in real world environments. Performance may also be affected by poor image quality and limited network connectivity.

Future work will focus on improving the framework in several areas. A larger and more diverse dataset can help evaluate the system across more manufacturers, languages, packaging styles, and regions. Integration with authoritative regulatory and manufacturer databases would also allow license numbers, batch information, and other package details to be checked against trusted sources. Further improvements could include better image quality assessment, stronger uncertainty estimation, support for low bandwidth environments, and comparison with specialized computer vision and retrieval based verification approaches.

Overall, the results of this study show that improving the reliability of multimodal AI does not always require retraining a larger model. Careful system design and clear constraints on model behavior can also have a significant impact. By

treating the multimodal model as a structured verification component rather than an unrestricted conversational system, MedTruth was able to produce more consistent and actionable results. This approach provides a useful direction for applying multimodal AI to other safety critical tasks where reliable and repeatable decisions are more important than unrestricted generation.

VII. DECLARATIONS

Declaration	Statement
Corresponding Author	Dr. Pankaj Kunekar
Funding	The authors declare that no funding was received for the research, authorship, or publication of this manuscript.
Consent to Publish	Not applicable.
Ethics and Consent to Participate	Not applicable.
Data Availability	The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

REFERENCES

- [1] K. Singhal et al., Large language models encode clinical knowledge, *Nature*, vol. 620, pp. 172 to 180, 2023.
- [2] A. J. Thirunavukarasu et al., Large language models in medicine, *Nature Medicine*, vol. 29, pp. 1930 to 1940, 2023.
- [3] Z. Ji et al., Survey of hallucination in natural language generation, *ACM Computing Surveys*, vol. 55, no. 12, 2023.
- [4] H. Nori et al., Capabilities of GPT-4 on medical challenge problems, *arXiv preprint arXiv 2303.13375*, 2023.
- [5] M. Agrawal et al., Large language models are zero shot clinical information extractors, *arXiv preprint arXiv 2205.12689*, 2022.
- [6] M. Bansal et al., Deep learning techniques for counterfeit drug detection, *IEEE International Conference on Computing*, 2021.
- [7] R. Kumar and R. Tripathi, Traceability of counterfeit drugs using blockchain, *Wireless Personal Communications*, vol. 110, 2019.
- [8] T. K. Mackey et al., The global counterfeit drug trade patient safety and public health risks, *Journal of Pharmaceutical Sciences*, vol. 104, 2015.
- [9] C. Picot et al., Mobile technologies for counterfeit drug detection, *PLoS One*, vol. 7, 2012.
- [10] Y. Ren et al., Smartphone based spectrometry for drug screening, *IEEE Sensors Journal*, vol. 20, 2020.
- [11] A. Radford et al., Learning transferrable visual models from natural language supervision, *ICML*, 2021.
- [12] J. Alayrac et al., Flamingo a visual language model for few shot learning, *NeurIPS*, 2022.
- [13] Gemini Team, Gemini a family of highly capable multimodal models, *Google DeepMind*, 2023.
- [14] H. Liu et al., Visual instruction tuning, *NeurIPS*, 2023.
- [15] Z. Zhang et al., Multimodal chain of thought reasoning in language models, *arXiv preprint arXiv 2302.00923*, 2023.
- [16] J. Wei et al., Chain of thought prompting elicits reasoning in large language models, *NeurIPS*, 2022.
- [17] J. White et al., A prompt pattern catalog to enhance prompt engineering with ChatGPT, *arXiv preprint arXiv 2302.11382*, 2023.
- [18] S. Mishra et al., Reframing instructional prompts to GPT-3 calls, *ACL*, 2022.
- [19] T. Heston and C. Khun, Prompt engineering in medical education, *BMC Medical Education*, 2023.
- [20] J. Wang et al., On the robustness of ChatGPT an adversarial and out of distribution perspective, *arXiv preprint arXiv 2302.12095*, 2023.