STANDARDIZING INTELLIGENCE: Aligning Generative AI for Regulatory and Operational Compliance

Joseph Marvin Imperial[1,2] **Matthew D. Jones**[3] **Harish Tayyar Madabushi**[1,2]

[1]UKRI CDT in Accountable, Responsible and Transparent AI
[2]Department of Computer Science [3]Department of Life Sciences
University of Bath, UK

jmri20@bath.ac.uk prpmdj@bath.ac.uk htm43@bath.ac.uk

Abstract

Technical standards, or simply *standards*, are established documented guidelines and rules that facilitate the interoperability, quality, and accuracy of systems and processes. In recent years, we have witnessed an emerging *paradigm shift* where the adoption of generative AI (GenAI) models has increased tremendously, spreading implementation interests across standard-driven industries, including engineering, legal, healthcare, and education. In this paper, we assess the *criticality levels* of different standards across domains and sectors and complement them by grading the current *compliance capabilities* of state-of-the-art GenAI models. To support the discussion, we outline possible challenges and opportunities with integrating GenAI for standard compliance tasks while also providing actionable recommendations for entities involved with developing and using standards. Overall, we argue that *aligning GenAI with standards through computational methods can help strengthen regulatory and operational compliance*. We anticipate this area of research will play a central role in the management, oversight, and trustworthiness of larger, more powerful GenAI-based systems in the near future.

1 Introduction

The industries of the modern world rely on systematic processes for the efficient production of goods and delivering services guided through **standards**. According to the International Organization for Standardization (ISO)[1], standards refer to a general form of documented specifications, rules, and norms specialized across various domains and sectors such as healthcare, education, engineering, science, communication, security and defense. For example, in the aerospace engineering domain, any technical instruction manual produced must conform to recognized technical language standards such as ASD-STE100 Simplified Technical English (STE)[2] developed and maintained by the Aerospace, Security and Defence Industry Association of Europe (ASD, formerly AECMA). In the education and language proficiency assessment domain, on the other hand, teachers in charge of developing

[1]https://www.iso.org/standards.html
[2]https://www.asd-ste100.org/

First published arXiv:2503.04736v1 3 Feb 2025

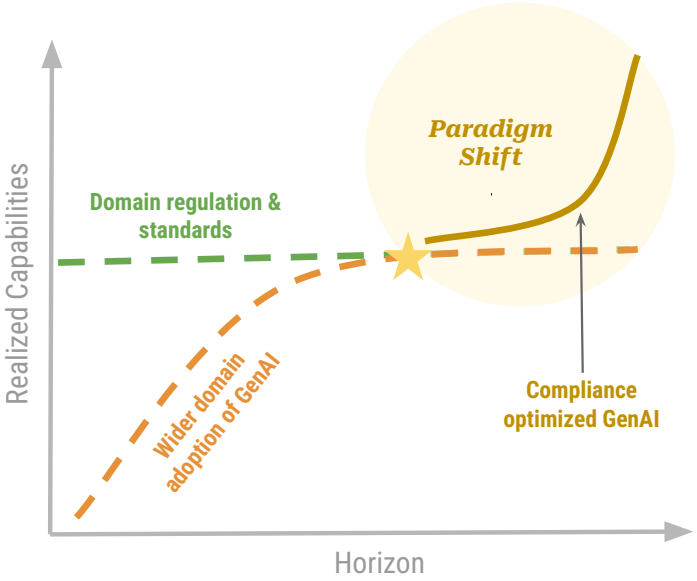

Figure 1: We describe an emerging **paradigm shift** where domain experts from interdisciplinary areas such as education, engineering, and healthcare are using advanced generative AI models (e.g., GPT-4) to assist them with regulatory and operational compliance through standards. We pattern the temporal observation of the paradigm shift within the near-to-midterm realized capabilities of GenAI as described in Eiras et al. [29].

curriculum materials must follow education standard frameworks such as the Common European Framework of Reference for Languages (CEFR)[3] in Europe or the Common Core State Standards (CCS)[4] in North America to produce high-quality classroom content.

Standards are typically composed of carefully defined technical *specifications* for measurements, design, and performance quotas that can be used to check or validate regulatory and operational compliance while preserving interoperability, quality, and accuracy [6]. In recent years, there has been observable interest from users across various domains in adopting these instruction-following generative AI (GenAI) technologies, such as ChatGPT, due to their documented capabilities, including the ability to follow complex instructions and generate human-like writing. For example, a survey by the Department of Education revealed that 62% of primary and secondary teachers in the UK have reported using GenAI tools to create new educational content and lessons for classroom use[5] [25]. Complementary to this, recent empirical works on benchmarking GenAI models for automatic content generation using CEFR and CCS standards as references for control show that off-the-shelf commercial and open-weight LLMs such as Llama2 [96] and GPT-4 [1] can be systematically steered to produce high-quality content through methods such as in-context learning (ICL) and reinforcement learning (RL) while preserving high automatic and human expert evaluations [45, 63]. Thus, this promising research direction of aligning GenAI models with standards underscores the need for greater attention from both the AI and interdisciplinary research communities to examine how GenAI is transforming regulatory and operational compliance across standard-driven domains.

In this paper, we analyze the changing landscape—an emerging *paradigm shift*—of regulatory and operational compliance through standards across various sectors and domains. We propose a joint **CRITICALITY AND COMPLIANCE CAPABILITIES FRAMEWORK (C3F)** for assessing the measured capabilities of 15 recent and community-recognized foundational and specialized GenAI models for standard-compliance tasks and the criticality levels of 34 standards from various domains based on their sensitivity and potential consequences in case of non-compliance. We cover a variety of case studies supporting the paradigm shift, including healthcare, education, safety, finance, and

[3] https://www.coe.int/en/web/common-european-framework-reference-languages

[4] https://www.thecorestandards.org/read-the-standards/

[5] n = 230. Data gathering was conducted around August and September 2023 with respondents from 23 educational settings.

engineering. We outline possible challenges and opportunities with learning standard compliance using GenAI models, the benefits if done successfully, and recommendations for various stakeholders involved. Finally, we take the following position that **aligning GenAI with standards through computational methods can help strengthen regulatory and operational compliance**. Thus, leading to enhanced control, oversight, and trust among these systems in real-world settings.

2 The Regulatory and Operational Compliance Landscape

To set the stage, we discuss preliminary information on the common definitions and processes comprising the development of standards, including common characteristics and entities involved in their conception and development.

2.1 Defining A Standard

Standards are established to provide specifications for measuring quality and proof of compliance with regulations. In line with this, we consider standards as *expert-defined* documents since one typically needs substantial knowledge within a domain or sector to propose measurable requirements and regulatory orders for compliance from users, industries, and organizations [24]. The typology of standards can be generalized into either *product-based* standards that specify target characteristics for physical or digital products or *non-product* standards that govern and specify target efficiency, operation, and performance-based measures for processes and services [72].

To cater to the diverse overlapping notions of standards, in this paper, we do not restrict the scope of standards to those created by international and regional standard developing organizations (SDOs) such as ISO, IEEE, ETSI, CEN-CENELEC, or NIST. We also include formally documented rules coined under related terms like *frameworks*, *guidelines*, and *checklists*, which are often used in interdisciplinary domains as forms of standards themselves since they observe objectively similar nature and usage. In summary, to unify the overlapping characteristics, a documented set of guidelines can be considered a standard if it conforms to the following below:

Anatomy: Standards are composed of well-defined specifications and procedures documenting measurable requirements for a given product or process.

Purpose: Standards serve specific purposes for various domains and sectors, such as introducing a formal language of communication, defining recognized procedures, ensuring compatibility checks, specifying performance requirements, and assuring compliance with regulations.

Recognition: Standards are developed and recognized by members and constituents of a private or public domain, sector, organization, or regulatory body.

2.2 Standards as Co-Regulation and Co-Integration Tools

A standard may be developed publicly or privately through initiatives by the government or regulatory bodies, unions, organizations, and expert groups. In legislation, a standard may be paired with specific laws as a form of a **co-regulation tool** which, if successfully fulfilled, can serve as a *form of compliance with related state or nation-wide jurisdictions* [78]. In this case, a regulatory body may appoint one or more SDOs to initiate the workflow of creating a specific standard that contains the legal requirements that must be included in line with the law. An example of these standards includes the well-known ISO/IEC 27001:2022 which defines measures for Personally Identifiable Information (PII) controllers and processors of any information security management system (ISMS) in response to the EU General Data Protection Regulation (GDPR) for data privacy and protection [48].

On the other hand, standards developed through non-legislative initiatives can serve as a **co-integration tool** which focuses on organization- or community-wide *interoperability and harmonization of systems and processes*. A well-known example is the Web Standards developed by the World Wide Web Consortium (W3C)[6] which maintains all technical specifications, guidelines, and protocols for web-based technologies including HTML, CSS, and XML. These standards and technologies are globally recognized and form the core building blocks of the Web or Internet.

[6] https://www.w3.org/standards/

3 Paradigm Shift with GenAI

A **paradigm shift** occurs when a dominant standard practice becomes incompatible due to some emerging technological phenomena, facilitating the adoption of new forms of conceptualization, practices, or paradigms [54, 94]. The current state-of-the-art GenAI models are known to exhibit remarkable capabilities across a wide range of generative tasks. In particular, one of the most useful and powerful skills a model can learn is the ability to *follow complex human instructions* from prompts [18, 74, 99]. Standards are composed of technical specifications which, at their core, can also be considered a set of instructions. As such, it was not long until users and practitioners knowledgeable of standards in their specific domains and sectors started exploring and reframing these specifications as instructional prompts for GenAI models (e.g., GPT-4) to assist with compliance-based tasks. We consider this phenomenon as an *emerging paradigm shift* in standards and regulatory compliance, as shown in Figure 1. This paradigm shift is introduced by the rise of instruction and preference-optimized GenAI models that can follow specifications derived from standards through well-structured prompting techniques and domain-specific fine-tuning.

In this section, we further discuss specific cases from the literature in relation to the paradigm shift observed in two major aspects: 1) **conformity assessment** practices with standards and 2) **generating standard-aligned content** across various domains and sectors.

3.1 Shift in Standard Conformity Assessment

Conformity assessment, in relation to standards, pertains to how implementing organizations and users measure the level to which their products or services meet the requirements of the standard itself. This process is often formally known as *certification* and is extremely variable and dependent on several factors, including the level of conformity required by the standard and common assessment norms in specific domains or sectors [46]. For some standards-driven sectors such as pharmaceuticals, automotive, and energy industries, certifications are a legal or contractual requirement. For non-regulatory and operational standards, certifications are less common, and conformity can often be assessed through various means, including using third-party evaluator software, hiring trained expert evaluators, or self-assessment by learning the standards from publications or documentation releases. We highlight notable works across various domains in automating conformity assessment with GenAI below:

Data Privacy Laws. Well-known data privacy laws such as the General Data Protection Regulation (GDPR)[7] [81] for the EU and the Health Insurance Portability and Accountability Act (HIPAA)[8] [2] for the US have served as common ground for optimizing GenAI models in terms of compliance checking due to the availability of data. Fan et al. [31] proposed the GoldCoin framework, which leverages contextual integrity theory to build synthetic case scenarios showing compliance and violations of the HIPAA Privacy Rule. The works of Zoubi et al. [108] and Zhu et al. [106] introduced PrivT5 and LegiLM, new specialized finetuned models trained from compilations of GDPR-related legal content such as case laws and data-sharing contracts and reported state-of-the-art performance in legal compliance tasks in NLP.

Financial and Accounting Report Standards. The International Financial Reporting Standards (IFRS) provides a standardized method of evaluating a company's financial performance for compliance across national and international regulations [76]. Auditing financial documents for compliance is considered labor-intensive, and the exploration of AI-driven solutions has been evident in recent years [4]. The work of Berger et al. [8] in collaboration with PwC Germany reported the effectiveness of GPT-4 for compliance validation of text sections from financial reports concerning IFRS and Germany's Handelsgesetzbuch or Commercial Code (HGB) through template-based prompting while noting the need for a major improvement in robustness before deploying to real-world scenarios.

Operational Design Standards for Driving Autonomous Systems. Beyond text-based applications, GenAI models have also been explored and have shown promising results for compliance assessment in multimodal settings. The work of Hildebrandt et al. [40] examined the use of OpenAI's ChatGPT-4V [73] and Vicuna [15] integrated in a pipeline called ODD-diLLMma to check the compliance of compiled sensor image data from self-driving cars with respect to Operational Design Domains

[7] https://gdpr-info.eu/
[8] https://www.hhs.gov/hipaa/index.html

(ODDs). ODDs are documented standards provided by manufacturers (e.g., Tesla or GMC) describing specific conditions under which a self-driving car may operate safely and within its designed function (e.g., *vehicle must not be driven at night*). The proposed ODD-DILLMMA pipeline is considered the first to automate the compliance checking of ODDs using multimodal LLMs with high accuracy between $\tilde{8}5\%$-94% across 11 weather, environment, and roadway characteristic dimensions.

3.2 Shift in Standard-Aligned Content Generation

Automatically generating text- or image-based content that adheres to a specific set of detailed specifications is considered a challenging task, even for AI-based models. State-of-the-art language models like GPT-4 can be prompted to create content using seed topics in which specific syntactic and semantic characteristics can be directed [60, 105]. Vision-based and multimodal models have also demonstrated the same level of controllability through prompting, particularly in models like DALL-E [9] and Stable Diffusion [83]. This degree of controllability via simple interactions through a chat interface, which can easily be utilized by users, has been pivotal to the applications of these models across various domains and sectors. We emphasize previous studies that focused on improving GenAI models' capabilities to automatically generate content that conforms to the standards below:

Education and Language Proficiency Frameworks. Content-based standards serve as a meter to ensure that classroom resources, such as reading and activity books, meet certain research-based quality criteria [55, 84]. An example of a content standard is the Common European Framework of Reference for Languages (CEFR), which is one of the most used resources for automatic educational content generation tasks using LLMs. The combined recent works of Imperial et al. [45], Malik et al. [63], and Glandorf and Meurers [33] have all explored a wide range of LLMs, including Llama, GPT-4, and Mistral, using specifications from CEFR in prompts to steer for desired granularities, including target complexity, grammar rules and structure, and levels of meaning. Experts in language testing using CEFR in Imperial et al. [45] have also given positive feedback on how GPT-4 can achieve a certain level of completeness, fluency, and coherence in generated texts.

Medical Reporting and Appraisal Standards. High-quality documentation and appraisal in the medical literature are driven by checklists and reporting standards. Sanmarchi et al. [86] explored ChatGPT's capabilities to reformulate the STROBE checklist [97] to analyze epidemiology studies related to COVID-19 vaccinations in 68 countries. The results of the study support ChatGPT's potential as an assistant in setting up epidemiological observational research but caution against its tendency to produce inconsistent responses when analyzing methods. In the same vein, Muluk [71] also used ChatGPT for customizing checklists related to managing patient-specific musculoskeletal injections that also conform to the METRICS standard [85]. The study echoed the considerable potential of GenAI models like ChatGPT for streamlining easily verifiable aspects in clinical practices, such as preparing medical checklists, but emphasized the importance of expert oversight.

Industry Safety Policies. Recent works have explored integrating industry safety policies into GenAI models to improve their capability to generate content that adheres to these specifications. An example of this is the Deliberative Alignment training paradigm used in OpenAI's O-series model [34]. In this method, LLMs are optimized to interpret the company's safety specifications and reason over them when responding to potentially harmful prompts. Another advantage of this method is that the models have been optimized to identify which policy specifications might be relevant to produce a compliant response, rather than going through the full copy of the policy at every iteration. Likewise, the work of Zhang et al. [102] also observes a similar approach to safety policy alignment but focuses on controlling the levels of safety by retraining models across different providers (e.g., *safety policies for generating realistic dialogues for video games can be relaxed to allow cursing*).

4 CRITICALITY AND COMPLIANCE CAPABILITIES FRAMEWORK (C3F)

The level of sensitivity and compliance requirements vary by standard and depend on its purpose and scope (see discussion on compliance in Section 3). For example, coding standards such as Python Enhancement Proposals (PEP) are less sensitive and critical than healthcare standards like the HIPAA Privacy Rule. Thus, for researchers exploring areas using GenAI as an assistant with compliance-based tasks, it is important to know *both* the documented capabilities of the GenAI models they plan to use for experiments *and* the target level of standard compliance they use as a benchmark for success or failure. To bridge this gap, we propose C3F, a joint CRITICALITY

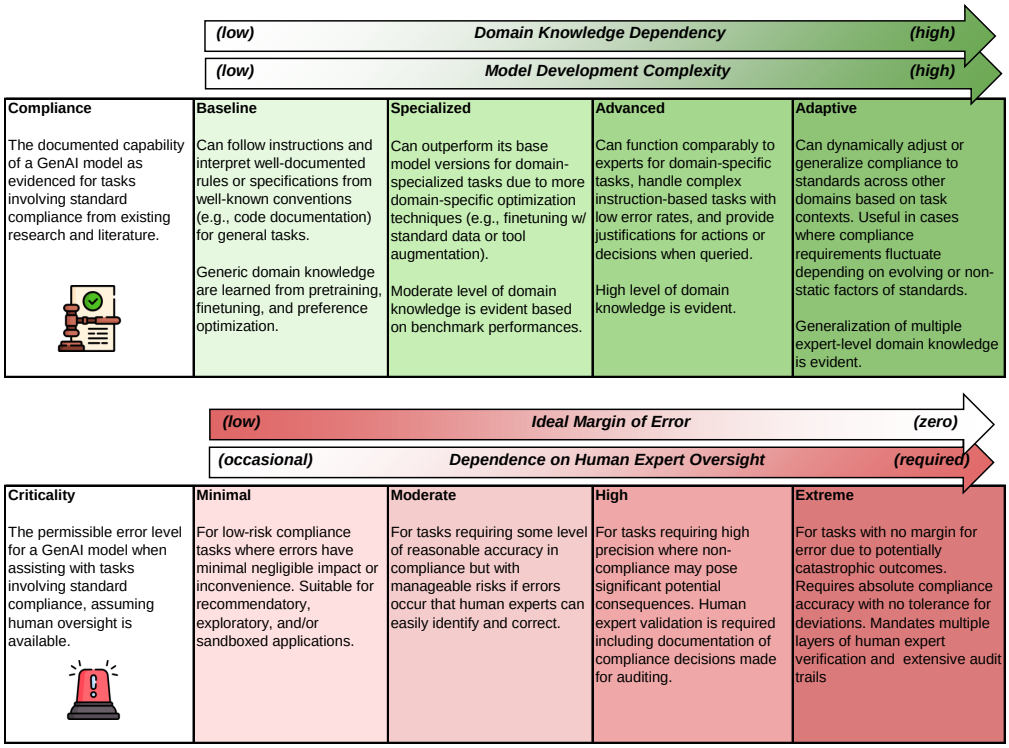

Figure 2: **The CRITICALITY AND COMPLIANCE CAPABILITIES FRAMEWORK (C3F).** We introduce a joint framework for assessing the current state-of-the-art foundational and specialized text and image-based GenAI models based on their (**Top**) documented *compliance capabilities* for generating content that aligns with standards, as well as (**Bottom**) the estimated *criticality* of standards from various domains and sectors based on the permissible error level a GenAI model can commit and the potential consequences in the case of non-compliance.

AND COMPLIANCE CAPABILITIES FRAMEWORK in Figure 2 to classify the current capabilities of modern GenAI models to follow compliance with standards as well as assess the different levels of criticality of standards across domains and sectors. We provide a more in-depth discussion of the two components of C3F in the succeeding sections.

4.1 Classification of GenAI by Documented Compliance Capabilities

We define **compliance capabilities** as an aggregation of a GenAI model's documented capabilities for compliance-based tasks across various publications and recognition from the interdisciplinary community. For C3F, we propose a four-level assessment scheme to measure a GenAI model's compliance capabilities in an increasing linear fashion as seen in Figure 2.

Compliance Capabilities Grading Scheme. For the Baseline level, we consider GenAI models that have been instruction-tuned as a *minimum qualification skill* for assessing compliance capabilities since the core nature of standards is to conform to their specifications. Instruction-tuned GenAI models (particularly LLMs) can pick up generic domain knowledge from the massive datasets often used for pretraining and additional output optimizations through instruction tuning and preference optimization. This is particularly evident from tasks such as code generation [7, 90] and health-related checklists [71, 86]. Succeeding levels, including Specialized and Advanced, require further evidence of domain knowledge expertise that outperforms Baseline-level models. GenAI models classified as Specialized are typically those that have been additionally fine-tuned with domain-specific gold-standard datasets such as clinical guidelines [14] and examples of regulatory compliant documents [31]. On the other hand, Advanced GenAI models are those that can be considered equal to domain experts for the task of standard compliance while also being capable of justifying or reasoning over

decisions with constraints from standards. Lastly, we consider Adaptive as the final level a GenAI model can obtain, which should exhibit the highest form of capability via generalization of multiple expert-level knowledge across domains. No existing GenAI model is currently classified under this level.

Assessment. We used the compliance capabilities component of C3F to assess 15 foundational and specialized GenAI models both for text and image as shown in Table 1 in Appendix C which includes information on their respective domains (in the case of Specialized models) and accessibility. We note that only the o-series models (o1 and o3) from OpenAI have been documented to exhibit the required capabilities to be classified under the Advanced category, as evidenced by their Deliberative Alignment study, which shows how LLMs can be trained to reason and select which applicable specifications from safety policies should be used to generate a safe response [34]. Recently released models, such as DeepSeek-R1 [35], are currently classified as Baseline and can be updated upon publication of literature documenting if their compliance capabilities can quality for the Advanced level.

Observation. We highlight two complementing observational points—(1) domain knowledge dependency and (2) model development complexity—in the compliance capabilities of the framework (reflected as two green gradient arrows in Figure 2). The first point, *domain knowledge dependency*, describes a direct relationship between the documented compliance capabilities of GenAI models and their evidenced domain knowledge. This is a straightforward observation as Specialized (and higher level) models will typically outperform their Baseline versions for domain-specific compliance tasks. The second point, *model development complexity*, reflects a similar direct relationship where higher compliance capabilities of GenAI models demands increasing complexity and costly data curation and training procedures. As a consequence, AI-based companies with higher financial resources such as Google, Meta, and OpenAI typically lead the development of more compute-heavy models.

4.2 Classification of Standards by Criticality Levels

We define the **criticality levels** of standards as a measure of their sensitivity, which can be determined by the allowable margin of error for a hypothetical GenAI model assisting with standard compliance tasks. For C3F, in parallel with compliance capabilities, we also propose a similar four-level assessment scheme illustrating the reduction of allowable errors as the criticality of standards increases as shown in Figure 2.

Criticality Level Grading Scheme. We consider Minimal criticality level as the least sensitive and can be used for GenAI-based experiments without requiring in-depth domain knowledge or expert oversight. Non-compliance can also be easily detected with existing rule-based software. This includes standards such as coding conventions (e.g., Python Enhancement Proposals) or writing and formatting guidelines (e.g., Plain Language, SMILES in Chemistry). For Moderate, there are potential risks associated with non-compliance but they can easily be managed and corrected by human experts. Examples in this category include most non-regulatory standards, standards with variations across domains, and standards developed by independent, private-sector organizations primarily used for interoperability, such as PRISMA for reporting systematic review papers and IFRS Accounting Standards in finance. Standards classified under High are those that require high levels of accuracy and may pose significant consequences in case of non-compliance. Most patient-facing healthcare standards classified in this category include the SPIRIT Checklist, which is used for transparency of clinical trial protocols; GDPR and HIPAA for data protection and privacy; and the USDA Food Safety Documentation. Lastly, the highest criticality level a standard can be classified as is Extreme, which is reserved for situations with zero margin of error allowed, and non-compliance may result in catastrophic and potentially irreversible consequences. This includes standards under the chemical, biological, radiological, and nuclear (CBRN) umbrella, such as the Safety Standards developed by the International Atomic Energy Agency and the Joint Operating Principles for Emergency Services, which pertain to documenting CBRN-related emergency responses. Such a high degree of sensitivity and criticality is necessary to include, as works on GenAI, particularly LLMs, are already gaining research attention and preliminary works [23, 41].

Assessment. We used the standard criticality level component of C3F to assess 34 globally recognized standards (including guidelines, checklists, and policies) from a wide range of domains and sectors. While the level of compliance with a certain standard can be deduced by reading its

respective documentation and release reports, collecting the opinions of domain experts can justify its classification based on criticality from C3F, which is a normal practice in the conventional standards development process shown in Figure 3 in Appendix C. In assessing the standards, we consider two things: the *consequences of harm* and the *scale of harm*. The former describes the level of potential damage that non-compliance with standards can trigger, while the latter considers the number of people who might be harmed by an error caused by non-compliance. For example, an error in a patient-facing scenario in healthcare might endanger one person, but an error in a nuclear energy scenario might endanger or kill thousands. Both can lead potentially to death, but the scale varies between the two. For standards classified under the domains of healthcare and engineering in Table 2, we conversed with two practitioners from our university network who have experience using the standard and obtained their assessments based on C3F.

Observation. We also highlight two observational points—(1) ideal margin of error and (2) dependence on human expert oversight—in terms of criticality levels proposed in the framework (reflected as two red gradient lines). In this case, however, the two points are opposites. The *ideal margin of error* should decrease from low to zero as criticality levels increase, particularly for standards rated High and Extreme, to avoid the potential consequences of non-compliance. The *dependence on human expert oversight*, on the other hand, is directly proportional and should increase from occasional (for Minimal criticality) to required (for High and Extreme) as criticality levels also increase.

5 Challenges and Opportunities

To complement the categorization of standards by criticality levels in Section 4, we provide an extensive discussion of the technical aspects that make the process of aligning GenAI models with standards both challenging and promising, and highlight the advantages that standards can introduce to existing GenAI-based systems

5.1 Complexities of Standards

Standards Are Living Documents Standards can be considered as *living documents* where their contents can be revised, changed, or updated at any point in time by their developers. SDOs typically perform periodic reviews of their standards (5 in the case of ISO), which are driven by factors such as industry changes, the introduction of new methods or products, and significant technological advances in the field [12, 27]. Other forms of standards we consider in this work, such as organizational policies and guidelines, can be updated as needed. As such, architectures for standard-aligned GenAI model pipelines should be designed to adapt dynamically, accommodating both minor and major changes in a standard's specifications for compliance-based tasks. A good example of this is ISO/IEC 22989:2022, where new AI-related terminologies and concepts (e.g., *AI agent*) are added to the list, including their recognized definitions as research in the field progresses [47].

Standards Are Specifications-Driven The core DNA of standards is its technical specifications, which detail precise descriptions of the inputs and outputs of products being measured and the expected behaviors of processes. This approach parallels established engineering disciplines, where clear, non-ambiguous specifications enable the development of modular and robust systems [93, 98]. The quantity of specifications that make up a complete standard depends on the standard's complexity. The challenge for GenAI models is to learn these specifications and apply them dynamically or on an as-needed basis. Moreover, showing which specifications have been used by the GenAI model to generate content or check for compliance can fulfill regulatory demands for transparency and improve user trust [28, 61, 100].

Standards Have Limited Reference Data The clarity and level of information needed by domain experts in interpreting input and output requirements and learning from a few standard-compliant examples may not be equivalent to what GenAI models require. This limitation may pose challenges when automating compliance or conformity assessment using GenAI models. Some standards, especially those focusing on prescribed semantic and syntactic information, such as MLCommons' AILuminate Standard for assessing safety responses of LLMs [69] or the NHS Standard for creating health-related content [26], may have specifications where only a few conforming examples (typically 2-3) are provided. For cases like these, researchers often need to conduct their own data collection from external resources or perform synthetic data generation to increase the gold-standard reference

data for compliance-based tasks, which may entail additional costs. These practices have been explored in previous works on using GenAI for compliance assessment concerning the HIPAA Privacy Rule and GDPR Documentation Standard [20, 31, 108].

Standards Depend on Domain Knowledge Expertise Standards developed by experts in fields such as healthcare, science, and engineering require domain knowledge to interpret how their specifications and constraints can be applied to domain-specific processes or activities that require compliance. Baseline GenAI models are often trained with massive collections of web-scraped internet data combined with scaling techniques from which it can learn generic, jack-of-all-trades knowledge required for most tasks [17, 18]. LLMs, in particular, often demonstrate better performance than their baseline counterparts in domain-specific tasks when further finetuned or optimized with additional curated datasets that provide deeper domain knowledge. [14]. This can be evidenced by previous works such as the Meditron-70B model, where they finetuned a Llama-70B with 40K clinical medical standards and guidelines from online healthcare websites and over 16M medical abstracts and papers from PubMed and PubCentral, which obtained higher performance across medical QA tasks than other closed and open-weight LLMs [14]. This static finetuning of large models may be effective only for well-established standards that will not be updated for a substantial amount of time but may not be suited for standards, policies, or guidelines that are inherently dynamic and flexible.

Standards Require Strong Expert-Level Evaluation Perhaps one of the most important aspects of aligning GenAI with standards for compliance tasks is how we evaluate such alignment in terms of accuracy and practical usability. As discussed in Section 3, conformity assessments with standards can be a variable process depending on the level of compliance required for inputs and on how protocols work across different domains and sectors. Thus, the process of evaluation for standard compliance tasks should anchor to an agile, use-case basis that targets specific output requirements rather than adopting a one-size-fits-all approach [52]. Ultimately, domain experts in standards should already be involved even at the conceptualization stage of GenAI-based workflows related to standard compliance to help identify plausible evaluation metrics to improve reliability and usability in real-world settings.

5.2 Advantages and Benefits of Aligning GenAI with Standards

Standard Alignment Can Enhance Quality and Interoperability GenAI models are often controlled through various experimental means, such as different forms of structured prompting [11, 89, 100], finetuning [17, 99, 103], and preference optimization [74, 79, 107], which have shown effectiveness across various tasks and domains. In line with this, standards can serve as a *reference of control* for these models to generate and refine their outputs based on the standard's specifications, as done in recent works on education and language proficiency assessment [45] and safe response generation using company policies [34]. Likewise, in relation to the emerging body of work with GenAI-based agents, aligning them with standards to produce an interconnected ecosystem can enable enhanced interoperability between inputs and outputs, thus improving efficiency, transparency, and the production of quality-controlled content.

Standard Alignment Can Improve Oversight, Transparency, and Auditing In high-stakes domains, human oversight of any AI-based system or interface is crucial for transparency and auditing [10, 51]. As such, controlling how GenAI models produce outputs by updating the specifications of standards and being able to trace back deviations through these changes will be extremely valuable in areas such as bias mitigation [32], fairness evaluations [95], and domain-specific tasks related to healthcare, finance, and legal decision-making [10, 66, 70]. Likewise, this form of oversight achieved by controlling with standards can also be scaled through *superalignment*[9] where a smaller GenAI teacher model specializing in a particular standard can regulate larger student models while exploiting its enhanced instruction-following capabilities [13, 36].

Standard Alignment Can Strengthen User Trust Building user trust is considered one of the most elusive challenges in the design of AI-based systems, as it can dictate the lifeline of how these systems will be adopted and used [53, 82, 88]. Bridging the same rules and regulations that domain experts follow, in the form of standards to control GenAI models, can potentially enhance process-based user trust. For example, a medical expert may feel much more confident in using a specialized open model like Meditron [14] to complete or assist with their tasks than in using

[9]First coined by OpenAI: https://openai.com/index/introducing-superalignment/

a black-box general-purpose model, simply from knowing that the former has undergone further training using massive collections of clinical guidelines and medical papers with which the expert is familiar. This level of transparency given to domain experts can be highly beneficial for earning and strengthening user trust, as it assures them of certainties in the performance expectations and design of standard-aligned GenAI models [53].

Standard Alignment Can Reduce Risk of Inaccuracies The results of the State of AI in 2024 survey conducted by McKinsey revealed that *inaccuracy*—the tendency of GenAI models to produce factually incorrect and unexpected results—as one of the highest risk factors hindering adoption across major organizations and industries. Such risks can cause a domino effect, including losing user trust, potential physical or mental harm, and financial losses for both consumers and businesses if not properly mitigated and controlled [65]. In line with this, the standard alignment of GenAI models can contribute to reducing inaccuracies by pairing it with architectural enhancements such as scaling and finetuning with massive collections of domain-specific data to improve domain knowledge, as evidenced in previous works on education [43, 45], legal [31, 106], and medicine [14]. Additionally, businesses can also use standard alignment as proof that their services conform to domain-specific regulations, thereby providing quality assurance to clients and consumers.

6 Risks, Responsibilities, and Recommendations

At this stage, we outline the potential risks and key responsibilities of each stakeholder group involved in the development of standards and advancement of GenAI and provide recommendations to align with the evolving practices in standard compliance.

For Government and Regulatory Bodies Regulations established by governing bodies can be considered one of the major drivers of standard development (see Section 2). However, two of the main concerns with regulations are the *risk of rigid standardization* which may compromise innovation efforts and hinder or slow down the beneficial applications of AI [3], and the *risk of regulatory gaps* due to vague stipulations and the unrealistic technical feasibility of legislations [77]. To address these, we recommend periodic *regulatory and policy adaptation* to revisit existing regulations and critically discuss how GenAI's evolving role impacts compliance practices with these regulations. Likewise, legislation aimed at developing a unified *meta-standard* can also be enacted to assist professional regulatory bodies in effectively updating their professional guidelines for specific workforces that will experience major job augmentations with the use of GenAI (e.g., physicians now using GenAI to support clinical decision-making for patients in the UK), in order to uphold ethical principles and ensure human oversight [38]. Lastly, since the challenge of regulatory compliance through standards is closely tied to advanced research on GenAI, we propose ongoing research funding for academic and industry partners to promote scientific advancements in accountability, responsibility, and transparency of GenAI.

For SDOs, Industry, and Academic Expert Groups Rethinking the conventional standard development process (see Figure 3) due to the shift in compliance practices with GenAI might be a major but necessary step for SDOs, industry associations, and academic expert groups. This is important to prevent the regulatory authority of technical standards from the *risk of being uninformative* for users wanting clear guidelines for standard compliance amid the progress of GenAI. We recommend establishing a dedicated committee responsible for updating and initiating maintenance revisions of existing published standards. These domain-specific committees can conduct their own studies and collaborate with other stakeholders to *closely monitor current practices, available tools, and limitations* in using GenAI for standard compliance [64]. Similarly, to boost advancements in GenAI research with standard compliance, we recommend open-sourcing machine-readable format of standards along with producing gold-standard compliant data for the research community.

For GenAI Researchers and Model Developers When an AI model is deployed in critical high-risk domains such as healthcare, legal, or engineering safety, it is the responsibility of researchers and developers to propose explainability techniques to interpret model outcomes. This aspect of GenAI research is important to avoid the *risk of eroding trust* among domain users who will use this technology. Thus, we recommend using standard compliance as an impactful case study of *explaining* the black-box nature of GenAI models. Since standards can be considered as co-regulation tools, developing novel approaches such as automatic audit trail generation for decisions made by standard-aligned GenAI and providing human-readable explanations is vital for its wider interdisciplinary

adoption [70, 91]. In addition, we also recommend that researchers collaborate with domain experts and explore using standards and regulatory documents as *references for control* for content generation tasks. This research direction will contribute to realistic applications of the capabilities of GenAI where the task of standard compliance with documents can be added to LLM benchmark evaluation suites such as HELM [59], ChatBot Arena [16], and BIG-Bench [92].

For Regulated Entities, Practitioners, and Users The final stakeholder group that will experience the greatest impact from GenAI are the regulated entities, practitioners, and users. Our most important recommendation is to practice and uphold the highest form of responsibility and accountability in using GenAI to comply with regulatory and operational requirements [19]. Professionals in regulated fields do not need to understand the full technical inner workings of GenAI. However, to reduce the *risk of misuse and over-reliance*, they should remain knowledgeable about the limitations of any GenAI model, including its tendency to produce inaccurate responses and exhibit limited domain expertise. Thus, we strongly recommend establishing a *solid, domain-specific foundation of AI literacy*, which is expected of an ideal professional who knows how to work with and maximize the potential of these intelligent tools. Finally, we recommend close collaboration with all stakeholders in the standard development process while providing feedback to enhance the overall usability and experience of standard compliance with GenAI models.

7 Alternative Views

While we emphasized and supported our position in the previous sections, we present two main alternative views that we consider equally valid and necessary to ensure a healthy discussion in this emerging new research direction.

Aligning GenAI for Compliance May Be Superficial GenAI can produce non-factual, hallucinated responses if it lacks sufficient domain knowledge for various tasks. Deploying GenAI-based assistants, particularly those classified below the Adaptive level in compliance capabilities in C3F, is not realistic since AI often struggles with *very specific edge cases and context-dependent ambiguities* that only experienced domain experts can realistically resolve. In response to this, we believe GenAI is *not* intended to take a significant portion of work from domain experts, but rather, to act as *first-line assistants* in managing automatic, repetitive preliminary tasks, thereby allowing domain experts to focus other parts of their work. We emphasize that maintaining *human oversight* is crucial for any AI-based workflow, as indicated in the critical levels assessment of standards in C3F. Using GenAI as first-line assistants is also viable for standard-compliance tasks requiring generating content that conforms to specific writing conventions (e.g., ASD-STE Simplified Technical English [5, 44]) and can be easily verified with rule-based software.

Aligning GenAI for Compliance Creates False Sense of Security Optimizing GenAI to follow rigid standards might result in users and practitioners being too *overreliant* on these systems. Even if users are domain experts themselves, they can subjected to a *false notion of security* from an over-confident model, which might lead to reduced human vigilance and possibly cause major to catastrophic errors from non-compliance. In response to this, we acknowledge that *trust is a multifaceted concept*, particularly for AI systems [75]. As discussed in Section 6, our position on aligning GenAI with standards entails the need for strong collaborative efforts across all stakeholders in designing *user-centric approaches*. This recommendation is crucial as it enables regulated entities, practitioners, and users to actively engage in the *process of standard alignment* for GenAI models, where they will understand the importance of human oversight and accountability in using these systems and the potential risks of overreliance.

8 Outlook

GenAI is transforming how we work and perform day-to-day tasks and will continue to do so in the near future. In this paper, we view such transformation as a *paradigm shift* and discuss how supporting this change can be considered *a fundamental step toward building trustworthy, controllable, and responsible GenAI systems* through standard alignment.

Based on our position, we make **two calls to action** across all stakeholder groups involved in the development of standards and GenAI:

1. We encourage all stakeholders to collectively recognize and adapt to the paradigm shift introduced by GenAI, which can serve as a powerful tool for assisting with compliance to regulations and standards, provided it is used responsibly and ethically.
2. We advocate for a global, unified effort among stakeholders to fully realize the potential of this research direction and ensure that GenAI remains a force for good in regulated domains.

We look forward to welcoming new research ideas, global partnerships, initiatives, and collaborations, whether they complement or contrast with the position we introduced, in line with the promising direction of aligning GenAI for regulatory and operational compliance through standards.

Acknowledgments and Disclosure of Funding

We are grateful to a number of people who contributed to the discussions on the position and provided constructive feedback on earlier versions of this paper: Orlando Chiarello (ASD-STE STEMG), Howard Benn (ETSI), Julian Padget (University of Bath, IEEE P7003 WG), James Davenport (University of Bath, BSI), Tory Frame, Matthew Hewitt, Marina De Vos, George Fletcher, and Jinha Yoon. JMI is supported by the National University Philippines and the UKRI Centre for Doctoral Training in Accountable, Responsible, and Transparent AI [EP/S023437/1] of the University of Bath.

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023. URL https://arxiv.org/abs/2303.08774.

[2] Act. Health Insurance Portability and Accountability Act of 1996. *Public law*, 104:191, 1996. URL http://www.eolusinc.com/pdf/hipaa.pdf.

[3] P. Aghion, A. Bergeaud, and J. Van Reenen. The impact of regulation on innovation. *American Economic Review*, 113(11):2894–2936, 2023. URL https://www.aeaweb.org/articles?id=10.1257/aer.20210107.

[4] F. Albuquerque and P. Gomes Dos Santos. Exploring ChatGPT's capabilities in solving accounting standards problems: the case of IAS 37. *Cogent Education*, 11(1):2412492, 2024. URL https://www.tandfonline.com/doi/pdf/10.1080/2331186X.2024.2412492#page=14.85.

[5] ASD-STE100. *ASD-STE100 Simplified Technical English*. Simplified Technical English Maintenance Group (STEMG), issue 9 edition, Jan. 2025. URL https://www.asd-ste100.org/.

[6] ASTM International. Form and Style for ASTM Standards, 2025. URL https://www.astm.org/form-style-for-astm-stds.html. Accessed: 2025-01-23.

[7] R. Beer, A. Feix, T. Guttzeit, T. Muras, V. Müller, M. Rauscher, F. Schäffler, and W. Löwe. Examination of Code generated by Large Language Models. *arXiv preprint arXiv:2408.16601*, 2024. URL https://arxiv.org/pdf/2408.16601.

[8] A. Berger, L. Hillebrand, D. Leonhard, T. Deuser, T. B. F. De Oliveira, T. Dilmaghani, M. Khaled, B. Kliem, R. Loitz, C. Bauckhage, et al. Towards Automated Regulatory Compliance Verification in Financial Auditing with Large Language Models. In *2023 IEEE International Conference on Big Data (BigData)*, pages 4626–4635. IEEE Computer Society, 2023. URL https://www.computer.org/csdl/pds/api/csdl/proceedings/download-article/1TUOuabhdXq/pdf.

[9] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, et al. Improving Image Generation with Better Captions. *Computer Science*, 2(3):8, 2023. URL https://cdn.openai.com/papers/dall-e-3.pdf.

[10] S. R. Bowman, J. Hyun, E. Perez, E. Chen, C. Pettit, S. Heiner, K. Lukošiūtė, A. Askell, A. Jones, A. Chen, et al. Measuring Progress on Scalable Oversight for Large Language Models. *arXiv preprint arXiv:2211.03540*, 2022. URL https://arxiv.org/abs/2211.03540.

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[12] BSI Group. Standards terminology: When is a standard no longer a standard?, 2024. URL https://knowledge.bsigroup.com/articles/standards-terminology-when-is-a-standard-no-longer-a-standard. Accessed: 2025-01-20.

[13] C. Burns, P. Izmailov, J. H. Kirchner, B. Baker, L. Gao, L. Aschenbrenner, Y. Chen, A. Ecoffet, M. Joglekar, J. Leike, et al. Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=ghNRg2mEgN.

[14] Z. Chen, A. H. Cano, A. Romanou, A. Bonnet, K. Matoba, F. Salvi, M. Pagliardini, S. Fan, A. Köpf, A. Mohtashami, et al. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. *arXiv preprint arXiv:2311.16079*, 2023. URL https://arxiv.org/abs/2311.16079.

[15] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

[16] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, B. Zhu, H. Zhang, M. Jordan, J. E. Gonzalez, et al. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. In *Forty-first International Conference on Machine Learning*, 2023. URL https://openreview.net/forum?id=3MW8GKNyzI.

[17] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. PaLM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023. URL https://www.jmlr.org/papers/v24/22-1144.html.

[18] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al. Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

[19] M. Coeckelbergh. Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and Engineering Ethics*, 26(4):2051–2068, 2020. URL https://link.springer.com/content/pdf/10.1007/s11948-019-00146-8.

[20] P. Colombo, T. Pires, M. Boudiaf, R. F. C. P. de Melo, G. Hautreux, E. Malaboeuf, J. Charpentier, D. Culver, and M. Desa. SaulLM-54B & SaulLM-141B: Scaling Up Domain Adaptation for the Legal Domain. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=NLUYZ4ZqNq.

[21] A. Creswell and M. Shanahan. Faithful Reasoning Using Large Language Models. *arXiv preprint arXiv:2208.14271*, 2022. URL https://arxiv.org/abs/2208.14271.

[22] I. da Cunha. Un redactor asistido para adaptar textos administrativos a lenguaje claro. *Procesamiento del Lenguaje Natural*, 69:39–49, 2022. URL http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6426.

[23] M. de Costa, M. Anwar, D. Lau, and I. Hammad. Classification of Safety Events at Nuclear Sites using Large Language Models. *arXiv preprint arXiv:2409.00091*, 2024. URL https://arxiv.org/pdf/2409.00091.

[24] D. Demortain. Standardising through concepts: The power of scientific experts in international standard-setting. *Science and Public Policy*, 35(6):391–402, 2008. URL https://academic.oup.com/spp/article/35/6/391/1673768.

[25] Department for Education. Generative AI in education: educator and expert views. Government report, Department for Education, 1 2024. URL https://www.gov.uk/government/publications/generative-ai-in-education-educator-and-expert-views.

[26] N. Digital. Standard for Creating Health Content, 2025. URL https://service-manual.nhs.uk/content/standard-for-creating-health-content. Accessed: 2025-01-21.

[27] EE Times. When Standards Change, 2018. URL https://www.eetimes.com/when-standards-change/. Accessed: 2025-01-20.

[28] U. Ehsan, P. Tambwekar, L. Chan, B. Harrison, and M. O. Riedl. Automated Rationale Generation: A Technique for Explainable AI and its Effects on Human Perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 263–274, 2019. URL https://dl.acm.org/doi/abs/10.1145/3301275.3302316.

[29] F. Eiras, A. Petrov, B. Vidgen, C. Schroeder De Witt, F. Pizzati, K. Elkins, S. Mukhopadhyay, A. Bibi, B. Csaba, F. Steibel, F. Barez, G. Smith, G. Guadagni, J. Chun, J. Cabot, J. M. Imperial, J. A. Nolazco-Flores, L. Landay, M. T. Jackson, P. Rottger, P. Torr, T. Darrell, Y. S. Lee, and J. N. Foerster. Position: Near to Mid-term Risks and Opportunities of Open-Source Generative AI. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 12348–12370. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/eiras24b.html.

[30] S. Elkins, E. Kochmar, J. C. Cheung, and I. Serban. How Teachers Can Use Large Language Models and Bloom's Taxonomy to Create Educational Quizzes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23084–23091, 2024. URL https://ojs.aaai.org/index.php/AAAI/article/download/30353/32395.

[31] W. Fan, H. Li, Z. Deng, W. Wang, and Y. Song. GoldCoin: Grounding large language models in privacy laws via contextual integrity theory. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3321–3343, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi:10.18653/v1/2024.emnlp-main.195. URL https://aclanthology.org/2024.emnlp-main.195/.

[32] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, and N. K. Ahmed. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3):1097–1179, 09 2024. ISSN 0891-2017. doi:10.1162/coli_a_00524. URL https://doi.org/10.1162/coli_a_00524.

[33] D. Glandorf and D. Meurers. Towards Fine-Grained Pedagogical Control over English Grammar Complexity in Educational Text Generation. In E. Kochmar, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, and Z. Yuan, editors, *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 299–308, Mexico City, Mexico, June 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.bea-1.24/.

[34] M. Y. Guan, M. Joglekar, E. Wallace, S. Jain, B. Barak, A. Heylar, R. Dias, A. Vallone, H. Ren, J. Wei, et al. Deliberative Alignment: Reasoning Enables Safer Language Models. *arXiv preprint arXiv:2412.16339*, 2024. URL https://arxiv.org/abs/2412.16339.

[35] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Ding, H. Xin, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Wang, J. Chen, J. Yuan, J. Qiu, J. Li, J. Cai, J. Ni, J. Liang, J. Chen, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia, M. Zhang, M. Zhang, M. Tang, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. Chen, R. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, S. Li, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

[36] J. Guo, H. Chen, C. Wang, K. Han, C. Xu, and Y. Wang. Vision Superalignment: Weak-to-Strong Generalization for Vision Foundation Models. *arXiv preprint arXiv:2402.03749*, 2024. URL https://arxiv.org/abs/2402.03749.

[37] S. Hao, T. Liu, Z. Wang, and Z. Hu. ToolkenGPT: Augmenting Frozen Language Models with Massive Tools via Tool Embeddings. *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 36:45870–45894, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8fd1a81c882cd45f64958da6284f4a3f-Abstract-Conference.html.

[38] Y. Hashem, S. Esnaashari, D. Morgan, J. Francis, A. Poletaev, F. Enock, and J. Bright. One in Four UK Doctors Are Using Artificial Intelligence: Exploring Doctors' Perspectives on AI After the Emergence of Large Language Models, 2024. URL https://www.turing.ac.uk/news/publications/one-four-uk-doctors-are-using-artificial-intelligence.

[39] J. Hernandez, D. Golpayegani, and D. Lewis. An Open Knowledge Graph-Based Approach for Mapping Concepts and Requirements between the EU AI Act and International Standards. *arXiv preprint arXiv:2408.11925*, 2024. URL https://arxiv.org/abs/2408.11925.

[40] C. Hildebrandt, T. Woodlief, and S. Elbaum. ODD-diLLMma: Driving Automation System ODD Compliance Checking using LLMs. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13809–13816. IEEE, 2024. URL https://carl-h.com/assets/files/publications/IROS24-ODD.pdf.

[41] K. Hirata, Y. Matsui, A. Yamada, T. Fujioka, M. Yanagawa, T. Nakaura, R. Ito, D. Ueda, S. Fujita, F. Tatsugami, et al. Generative AI and large language models in nuclear medicine: current status and future prospects. *Annals of Nuclear Medicine*, pages 1–12, 2024. URL https://link.springer.com/article/10.1007/s12149-024-01981-x.

[42] J. Huang and K. C.-C. Chang. Towards Reasoning in Large Language Models: A Survey. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada, July 2023. Association for Computational Linguistics. doi:10.18653/v1/2023.findings-acl.67. URL https://aclanthology.org/2023.findings-acl.67/.

[43] J. M. Imperial and H. Tayyar Madabushi. Flesch or fumble? evaluating readability standard alignment of instruction-tuned language models. In S. Gehrmann, A. Wang, J. Sedoc, E. Clark, K. Dhole, K. R. Chandu, E. Santus, and H. Sedghamiz, editors, *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 205–223, Singapore, Dec. 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.gem-1.18/.

[44] J. M. Imperial and H. Tayyar Madabushi. SpeciaLex: A Benchmark for In-Context Specialized Lexicon Learning. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 930–965, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi:10.18653/v1/2024.findings-emnlp.52. URL https://aclanthology.org/2024.findings-emnlp.52/.

[45] J. M. Imperial, G. Forey, and H. Tayyar Madabushi. Standardize: Aligning language models with expert-defined standards for content generation. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1573–1594, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi:10.18653/v1/2024.emnlp-main.94. URL https://aclanthology.org/2024.emnlp-main.94/.

[46] International Organization for Standardization. Conformity Assessment. https://www.iso.org/conformity-assessment.html. Accessed: 2025-01-14.

[47] International Organization for Standardization. Iso/iec 22989:2022 - information technology — artificial intelligence — artificial intelligence concepts and terminology, 2022. URL https://www.iso.org/standard/74296.html. Accessed: 2025-01-20.

[48] ISO/IEC 27001:2022. Information security, cybersecurity and privacy protection — information security management systems — requirements, 2022.

[49] H. Ivison, Y. Wang, J. Liu, Z. Wu, V. Pyatkin, N. Lambert, N. A. Smith, Y. Choi, and H. Hajishirzi. Unpacking DPO and PPO: Disentangling Best Practices for Learning from Preference Feedback. *arXiv preprint arXiv:2406.09279*, 2024. URL https://arxiv.org/abs/2406.09279.

[50] S. Joseph, L. Chen, J. Trienes, H. Göke, M. Coers, W. Xu, B. Wallace, and J. J. Li. FactPICO: Factuality Evaluation for Plain Language Summarization of Medical Evidence. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8437–8464, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi:10.18653/v1/2024.acl-long.459. URL https://aclanthology.org/2024.acl-long.459/.

[51] Z. Kenton, N. Y. Siegel, J. Kramar, J. Brown-Cohen, S. Albanie, J. Bulian, R. Agarwal, D. Lindner, Y. Tang, N. Goodman, et al. On scalable oversight with weak LLMs judging strong LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=O1fp9nVraj.

[52] F. Khanzada. Conformity Assessment: Relevance of Quality in the Age of Industry 4.0. In *Handbook of Quality System, Accreditation and Conformity Assessment*, pages 1–28. Springer, 2024. URL https://link.springer.com/referenceworkentry/10.1007/978-981-99-4637-2_1-1.

[53] R. F. Kizilcec. How Much Information? Effects of Transparency on Trust in an Algorithmic Interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2390–2395, 2016. URL https://dl.acm.org/doi/abs/10.1145/2858036.2858402.

[54] T. Kuhn. The Nature of Scientific Revolutions. *Chicago: University of Chicago*, 197(0), 1970.

[55] P. M. La Marca, D. Redfield, and P. C. Winter. State Standards and State Assessment Systems: A Guide to Alignment. Series on Standards and Assessments. *Non-Journal*, 2000. URL https://files.eric.ed.gov/fulltext/ED466497.pdf.

[56] H. Lee, S. Phatale, H. Mansoor, T. Mesnard, J. Ferret, K. R. Lu, C. Bishop, E. Hall, V. Carbune, A. Rastogi, et al. RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=uydQ2W41KO.

[57] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

[58] Z. Li, H. Zhu, Z. Lu, and M. Yin. Synthetic data generation with large language models for text classification: Potential and limitations. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore, Dec. 2023. Association for Computational Linguistics. doi:10.18653/v1/2023.emnlp-main.647. URL https://aclanthology.org/2023.emnlp-main.647/.

[59] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, et al. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=iO4LZibEqW.

[60] D. Liu and V. Demberg. ChatGPT vs human-authored text: Insights into controllable text summarization and sentence style transfer. In V. Padmakumar, G. Vallejo, and Y. Fu, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1–18, Toronto, Canada, July 2023. Association for Computational Linguistics. doi:10.18653/v1/2023.acl-srw.1. URL https://aclanthology.org/2023.acl-srw.1/.

[61] J. Liu, K. Marriott, T. Dwyer, and G. Tack. Increasing user trust in optimisation through feedback and interaction. *ACM Transactions on Computer-Human Interaction*, 29(5):1–34, 2023. URL https://dl.acm.org/doi/pdf/10.1145/3503461.

[62] A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, and P. Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, pages 1–11, 2024. URL https://www.nature.com/articles/s42256-024-00832-8.

[63] A. Malik, S. Mayhew, C. Piech, and K. Bicknell. From tarzan to Tolkien: Controlling the language proficiency level of LLMs for content generation. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15670–15693, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi:10.18653/v1/2024.findings-acl.926. URL https://aclanthology.org/2024.findings-acl.926/.

[64] D. Manheim, S. Martin, M. Bailey, M. Samin, and R. Greutzmacher. The Necessity of AI Audit Standards Boards. *arXiv preprint arXiv:2404.13060*, 2024. URL https://arxiv.org/pdf/2404.13060v1.

[65] McKinsey & Company. The State of AI in Early 2024: Gen AI Adoption Spikes and Starts to Generate Value. 2024. URL https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai. Accessed: 2025-01-22.

[66] B. Meskó and E. J. Topol. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digital Medicine*, 6(1):120, 2023. URL https://www.nature.com/articles/s41746-023-00873-0.

[67] L. J. V. Miranda, Y. Wang, Y. Elazar, S. Kumar, V. Pyatkin, F. Brahman, N. A. Smith, H. Hajishirzi, and P. Dasigi. Hybrid Preferences: Learning to Route Instances for Human vs. AI Feedback. *arXiv preprint arXiv:2410.19133*, 2024. URL https://arxiv.org/abs/2410.19133.

[68] S. Mishra, D. Khashabi, C. Baral, Y. Choi, and H. Hajishirzi. Reframing Instructional Prompts to GPTk's Language. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 589–612, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi:10.18653/v1/2022.findings-acl.50. URL https://aclanthology.org/2022.findings-acl.50/.

[69] MLCommons. AILuminate: A Collaborative, Transparent Approach to Safer AI, 2025. URL https://mlcommons.org/ailuminate/. Accessed: 2025-01-21.

[70] J. Mökander, J. Schuett, H. R. Kirk, and L. Floridi. Auditing Large Language Models: A Three-Layered Approach. *AI and Ethics*, pages 1–31, 2023. URL https://link.springer.com/article/10.1007/s43681-023-00289-2.

[71] S. Y. Muluk. Enhancing Musculoskeletal Injection Safety: Evaluating Checklists Generated by Artificial Intelligence and Revising the Preformed Checklist. *Cureus*, 16(5):e59708, 2024. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11150897/.

[72] National Science Foundation. Science and engineering indicators 2018, 2018. URL https://www.nsf.gov/statistics/2018/nsb20181/. Accessed: 2025-01-23.

[73] OpenAI. GPT-4V System Card, 2023. URL https://openai.com/index/gpt-4v-system-card/. Accessed: 2025-01-14.

[74] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

[75] A. Papenmeier, D. Kern, G. Englebienne, and C. Seifert. It's Complicated: The Relationship between User Trust, Model Accuracy and Explanations in AI. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 29(4):1–33, 2022. URL https://dl.acm.org/doi/full/10.1145/3495013.

[76] E. Posner. Sequence as Explanation: The International Politics of Accounting Standards. *Review of International Political Economy*, 17(4):639–664, 2010. URL https://scholar.google.com/scholar?output=instlink&q=info:_GvPJNuJ0xkJ:scholar.google.com/&hl=en&as_sdt=0,5&scillfp=7528166911765330717&oi=lle.

[77] H. Pouget. The EU's AI Act Is Barreling Toward AI Standards That Do Not Exist. *Lawfare*, 2023. URL https://www.lawfaremedia.org/article/eus-ai-act-barreling-toward-ai-standards-do-not-exist. Accessed: 2025-01-24.

[78] H. Pouget and R. Zuhdi. AI and Product Safety Standards under the EU AI Act, 2024. URL https://carnegieendowment.org/research/2024/03/ai-and-product-safety-standards-under-the-eu-ai-act?lang=en&center=middle-east. Accessed: 2025-01-07.

[79] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *Advances in Neural Information Processing Systems*, 36, 2024. URL https://dl.acm.org/doi/abs/10.5555/3666122.3668460.

[80] O. Ram, Y. Levine, I. Dalmedigos, D. Muhlgay, A. Shashua, K. Leyton-Brown, and Y. Shoham. In-Context Retrieval-Augmented Language Models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023. doi:10.1162/tacl_a_00605. URL https://aclanthology.org/2023.tacl-1.75/.

[81] P. Regulation. Regulation (EU) 2016/679 of the European Parliament and of the Council. *Regulation (EU)*, 679:2016, 2016.

[82] J. Riegelsberger, M. A. Sasse, and J. D. McCarthy. The Mechanics of Trust: A Framework for Research and Design. *International Journal of Human-Computer Studies*, 62(3):381–422, 2005. URL https://www.sciencedirect.com/science/article/pii/S1071581905000121.

[83] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. URL https://www.computer.org/csdl/proceedings-article/cvpr/2022/694600k0674/1H1iFsO7Zuw.

[84] D. R. Sadler. Academic achievement standards and quality assurance. *Quality in Higher Education*, 23(2):81–99, 2017. URL https://www.tandfonline.com/doi/pdf/10.1080/13538322.2017.1356614.

[85] M. Sallam, M. Barakat, M. Sallam, et al. A Preliminary Checklist (METRICS) to Standardize the Design and Reporting of Studies on Generative Artificial Intelligence–Based Models in Health Care Education and Practice: Development Study Involving a Literature Review. *Interactive Journal of Medical Research*, 13(1):e54704, 2024. URL https://pubmed.ncbi.nlm.nih.gov/38276872/.

[86] F. Sanmarchi, A. Bucci, A. G. Nuzzolese, G. Carullo, F. Toscano, N. Nante, and D. Golinelli. A step-by-step researcher's guide to the use of an AI-based transformer in epidemiology: an exploratory analysis of ChatGPT using the STROBE checklist for observational studies. *Journal of Public Health*, 32(9):1761–1796, 2024. URL https://link.springer.com/article/10.1007/s10389-023-01936-y.

[87] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom. Toolformer: Language Models Can Teach Themselves to Use Tools. *Advances in Neural Information Processing Systems*, 36:68539–68551, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html.

[88] P. Schmidt, F. Biessmann, and T. Teubner. Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, 29(4):260–278, 2020. URL https://www.tandfonline.com/doi/full/10.1080/12460125.2020.1819094.

[89] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online, Nov. 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.emnlp-main.346. URL https://aclanthology.org/2020.emnlp-main.346/.

[90] M. L. Siddiq, B. Casey, and J. Santos. A lightweight framework for high-quality code generation. *arXiv preprint arXiv:2307.08220*, 2023. URL https://arxiv.org/pdf/2307.08220.

[91] C. Song and V. Shmatikov. Auditing Data Provenance in Text-Generation Models. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 196–206, 2019. URL https://dl.acm.org/doi/abs/10.1145/3292500.3330885.

[92] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, et al. Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=uyTL5Bvosj.

[93] I. Stoica, M. Zaharia, J. Gonzalez, K. Goldberg, H. Zhang, A. Angelopoulos, S. G. Patil, L. Chen, W.-L. Chiang, and J. Q. Davis. Specifications: The missing link to making the development of LLM systems an engineering discipline. *arXiv preprint arXiv:2412.05299*, 2024. URL https://arxiv.org/abs/2412.05299.

[94] D. Tapscott and A. Caston. Paradigm Shift: The New Promise of Information Technology. *Economic Development Journal of Canada*, pages 62–66, 1994.

[95] C. Teo, M. Abdollahzadeh, and N.-M. M. Cheung. On Measuring Fairness in Generative Models. *Advances in Neural Information Processing Systems*, 36, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/220165f9c7f51163b73c8c7fff578b4e-Abstract-Conference.html.

[96] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open Foundation and Fine-Tuned Chat Modelss. *arXiv preprint arXiv:2307.09288*, 2023. URL https://arxiv.org/abs/2307.09288.

[97] E. Von Elm, D. G. Altman, M. Egger, S. J. Pocock, P. C. Gøtzsche, and J. P. Vandenbroucke. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *The Lancet*, 370(9596):1453–1457, 2007. URL https://www.thelancet.com/pdfs/journals/lancet/PIIS0140-6736(07)61602-X.pdf.

[98] H. Weber and H. Ehrig. Specification of modular systems. *IEEE Transactions on Software Engineering*, (7):784–798, 1986. URL https://www.computer.org/csdl/journal/ts/1986/07/06312979/13rRUyuNsyH.

[99] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022.

[100] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

[101] J. Ye, J. Gao, Q. Li, H. Xu, J. Feng, Z. Wu, T. Yu, and L. Kong. ZeroGen: Efficient Zero-shot Learning via Dataset Generation. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi:10.18653/v1/2022.emnlp-main.801. URL https://aclanthology.org/2022.emnlp-main.801/.

[102] J. Zhang, A. Elgohary, A. Magooda, D. Khashabi, and B. Van Durme. Controllable Safety Alignment: Inference-Time Adaptation to Diverse Safety Requirements. *arXiv preprint arXiv:2410.08968*, 2024. URL https://arxiv.org/pdf/2410.08968.

[103] L. Zhang, A. Rao, and M. Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. URL https://ieeexplore.ieee.org/abstract/document/10377881/.

[104] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. YU, S. Zhang, G. Ghosh, M. Lewis, L. Zettlemoyer, and O. Levy. LIMA: Less Is More for Alignment. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 55006–55021. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ac662d74829e4407ce1d126477f4a03a-Paper-Conference.pdf.

[105] W. Zhou, Y. E. Jiang, E. Wilcox, R. Cotterell, and M. Sachan. Controlled text generation with natural language instructions. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 42602–42613. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/zhou23g.html.

[106] L. Zhu, L. Yang, C. Li, S. Hu, L. Liu, and B. Yin. LegiLM: A Fine-Tuned Legal Language Model for Data Compliance. *arXiv preprint arXiv:2409.13721*, 2024. URL https://arxiv.org/pdf/2409.13721.

[107] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving. Fine-Tuning Language Models from Human Preferences. *arXiv preprint arXiv:1909.08593*, 2019. URL https://arxiv.org/abs/1909.08593.

[108] M. Zoubi, S. T.y.s.s, E. Rosas, and M. Grabmair. PrivaT5: A generative language model for privacy policies. In I. Habernal, S. Ghanavati, A. Ravichander, V. Jain, P. Thaine, T. Igamberdiev, N. Mireshghallah, and O. Feyisetan, editors, *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pages 159–169, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.privatenlp-1.16/.

A Hierarchy of the Standard Development Process

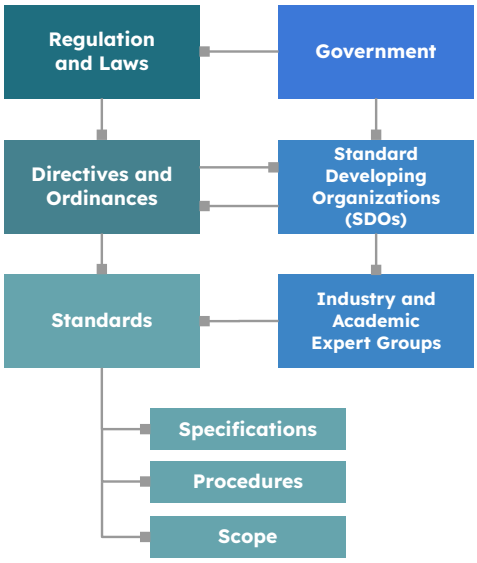

Figure 3: A supporting visualization of the general process of developing standards. Standards can be created either by government and regulatory bodies as a product of legislation or through industry associations and academic expert groups to ensure interoperability and quality of systems and processes.

B Technical Enhancements for Standard Alignment

We provide supplementary information on promising novel technical advancements that can be explored (but are not limited to) to further improve how GenAI models can align with standards and cater to their inherent complexities as discussed in Section 5. Some of these approaches may have been preliminarily explored by recent works for selected domains with available machine-readable standards data. Collaborations between AI and interdisciplinary areas can foster further advancements and open more novel approaches toward standard alignment.

Constraint and Knowledge Representations Specifications from standards can be represented as a form of a structured set of constraints either as input or as part of an embedded training regime for aligning GenAI models. The simplest example of this is through in-context learning (ICL), where constraints are framed as prompts in an instruction-like manner with specific informative examples provided to show the target output required from the model [11, 68]. This has been done on works such as the production of high-quality standard-conforming educational content with CEFR and Bloom's Taxonomy [30, 45, 63] and rewriting complex texts to conform to government-mandated plain languages guidelines [22, 50]. The advantage of ICL is its simplicity, which can easily be explored by non-technical domain users with any arbitrary GenAI-based chat interface (e.g., ChatGPT) and can be attributed to how the paradigm shift started (see Section 3). More advanced levels of representation focus on transforming standards into knowledge graphs and ontologies. An example of this is the work by Hernandez et al. [39], where they transformed the text content of the EU AI Act into a high-level knowledge graph to show links between defined terms and their associated requirements from the Act's statements for compliance checking.

Post-Training Improvements Post-training has become one of the major foci in ML research since the release of preference optimization techniques like RLHF [74] and DPO [79] combined with supervised finetuning (SFT) to improve the response alignment of GenAI models with respect to task requirements. All Baseline models in Table 1 have undergone post-training through SFT. For standard alignment, post-processing can potentially be emulated by aggregating prompt and response pairs that conform to the specifications of a standard and then finetuning a pretrained model with

this data. An example reference for this is OpenAI's Deliberative Alignment method [34] where an LLM is trained with chain-of-thought (CoT) style prompts [100] can identify whether which safety policy specification is applicable when identifying whether to respond to a prompt or not. However, for researchers who want to explore this approach, one caveat is that it will require at least 1,000 instances of very high-quality, expert-level pairwise data to achieve relatively decent performance [104]. Nonetheless, combinations of post-training techniques (e.g., SFT + DPO with standard-aligned preference pairs + CoT prompting) are viable approaches for improving a GenAI model's standard compliance capabilities.

Synthetic Data Generation Synthetic data generation using modern GenAI models that have undergone processes such as larger scaling, instruction-tuning, and preference optimization often outperform other data augmentation techniques for downstream tasks [58, 101]. In standard alignment of GenAI models, researchers can explore several options related to synthetic data generation to improve compliance capabilities. First, in parallel with post-training enhancements from above, using synthetic data in the form of standard compliant and non-compliant examples can be a practical choice to optimize a model's generation qualities. This approach has been applied by Fan et al. [31], where they generated synthetic case scenarios for GDPR and HIPAA Privacy Rules to finetune smaller LLMs for compliance detection. Moreover, recent works have documented higher performance for models that have been finetuned with a combination of high-quality expert data and machine-generated data in alignment tasks which makes the process of compiling standard-aligned feedback data relatively easier [49, 56, 67].

Retrieval and Tool Augmentation Augmenting GenAI models, particularly LLMs, with external tools to enhance their problem-solving capabilities has gained increasing research attention in recent years. The use of tools such as calculators, search engines, and API function calls has been shown to improve the zero-shot performance of LLMs across question-answering downstream tasks requiring up-to-date information [37, 87]. In the case of standard alignment, syntactic content-based standards such as the CEFR and CCS standards (see Table 2 for reference) requiring specific characteristics of texts such as sentence lengths to measure complexity can greatly benefit from a GenAI model that knows how to call a calculator tool to approximate how long sentences should be generated. Moreover, a search engine tool can also help GenAI models access updated versions of standard specifications from its original web sources as prior version checking. On the other hand, another approach that can improve GenAI models' domain knowledge is to encapsulate it in a retrieval-augmented generation (RAG) ecosystem where auxiliary retrievers that have access to external knowledge bases that can be added to as context to prompts [57, 80].

Reasoning Capabilities Whether GenAI models, such as LLMs, can reason or not is a highly debated topic in current ML research. Reasoning is an inherent ability that plays a crucial role in how humans solve problems through critical thinking [42]. Previous research often claims that such capability can be triggered in different ways, such as providing intermediary reasoning steps to prompts [100] or using arbitrary models to select and infer reasoning steps from context information [21]. Assuming GenAI models can actually reason, in standard alignment, such skill may play an important role in deciding which specifications of a standard are required to be followed and which ones can be disregarded safely, given the additional context information of a task. Preliminary work in this direction includes OpenAI's Deliberative Alignment method [34] where an LLM is optimized to reason over which safety policy specifications should be followed and which can be ignored using their o-series models. As such, GenAI models rated Advanced and Adaptive in C3F should document convincing reasoning capabilities across applicable tasks.

C Full Assessment Results of GenAI Models and Standards with C3F

As discussed in Section 4, we provide the full list of the 15 selected foundational and domain-finetuned GenAI models assessed based on their compliance capabilities and the 34 selected standards across multiple disciplines for their criticality levels.

MODEL	ORGANIZATION	DOMAIN	OPENNESS	COMPLIANCE
O-SERIES	OpenAI	General	Subscription	Advanced
GPT-4	OpenAI	General	Subscription	Specialized
DEEPSEEK-R1	DeepSeek-AI	General	Open Weight	Baseline
CLAUDE OPUS	Anthropic	General	Subscription	Baseline
GEMINI 2.0	Google	General	Subscription	Baseline
LLAMA 3.1 405B	Meta	General	Open Weight	Baseline
MISTRAL LARGE	Mistral	General	Subscription	Baseline
COMMAND-R 105B	Cohere	General	Open Weight	Baseline
MIDJOURNEY 6.1	Midjourney	General	Subscription	Baseline
DALL-E	OpenAI	General	Subscription	Baseline
MEDITRON 70B	Chen et al. [14]	Healthcare	Open Code	Specialized
GOLDCOIN-LLAMA	Fan et al. [31]	Healthcare, Legal	Open Code	Specialized
LEGILM	Zhu et al. [106]	Legal	Open Code	Specialized
CHEMCROW	M. Bran et al. [62]	Chemistry	Open Code	Specialized
STANDARDIZE-LLAMA	Imperial et al. [45]	Education	Open Code	Specialized

Table 1: We used the **compliance capabilities** component of C3F in Figure 2 to assess the latest versions of 15 foundational and specialized GenAI models both for text and image. We also include information on their respective domains (in the case of Specialized models) and accessibility (Open Weight, Open Code, or Subscription). The first section of the table includes industry-released GenAI models (mostly Baseline except for o-series models by OpenAI) while the second section is more focused on Specialized models commonly led by academic and research groups targeting specific domains and sectors. Models and their compliance capabilities in this table serve as a non-exhaustive example and can be extended or re-assessed as supporting literature are released.

STANDARD	DOMAIN	ORGANIZATION	OPENNESS	CRITICALITY
AILuminate Assessing Safety Standard	Software and Technology	MLCommons	Public	Minimal
Simplified Molecular Input Line Entry System (SMILES)	Chemistry	David Weininger	Public	Minimal
Associated Press Stylebook	Media and Communications	Associated Press	Subscription	Minimal
Plain Language Standards	Media and Communications	(region-specific)	Public	Minimal
Text Encoding Initiative	Software and Technology	TEI Consortium	Public	Minimal
ISO/IEC Systems and Software Engineering Documentation	Software and Technology	International Electrotechnical Commission	Subscription	Minimal
FAIR Data Principles Documentation	Software and Technology	GO FAIR Initiative	Public	Minimal

Continued on next page

STANDARD	DOMAIN	ORGANIZATION	OPENNESS	CRITICALITY
OpenAPI Specification (OAS)	Software and Technology	SmartBear Software	Public	Minimal
Section 508 Compliance Documentation	Software and Technology	US Government	Public	Minimal
Python Enhancement Proposals	Software and Technology	Python	Public	Minimal
Web Standards	Software and Technology	W3C	Public	Minimal
ASD-STE100 Simplified Technical English (STE)	Engineering	Aerospace, Security and Defence Industries Association of Europe	Public	Moderate
Common European Framework of Reference for Languages (CEFR)	Education	Council of Europe	Public	Moderate
Common Core Standards (CCS)	Education	National Governors Association, Council of Chief State School Officers	Public	Moderate
Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)	Healthcare	PRISMA Initiative	Public	Moderate
Standards for Quality Improvement Reporting Excellence (SQUIRE)	Healthcare	SQUIRE Initiative	Public	Moderate
IFRS Accounting Standards	Finance	International Financial Reports Standards	Public	Moderate
Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT)	Healthcare	National Health Service	Public	High
NHS Health Content Standards	Healthcare, Government	National Health Service	Public	High
HIPAA Privacy Rule	Healthcare, Government	US Government	Public	High
GDPR Documentation Standard	Legal, Government	European Union	Public	High
International Classification of Diseases (ICD) Standard	Healthcare, Government	Centers for Disease Control and Prevention	Public	High
Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Guidelines	Healthcare	STROBE Initiative	Public	High
Case Report Guidelines (CARE)	Healthcare	CARE Initiative	Public	High
Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT)	Healthcare	SPIRIT Initiative	Public	High
Digital Imaging and Communications in Medicine (DICOM)	Software and Technology, Healthcare	National Electrical Manufacturers Association	Public	High

Continued on next page

STANDARD	DOMAIN	ORGANIZATION	OPENNESS	CRITICALITY
USDA Food Safety Documentation	Healthcare, Government	US Government	Public	High
AGREE Reporting Checklist	Healthcare	International Appraisal of Guidelines, Research and Evaluation (AGREE)	Public	High
NICE Process and Methods	Healthcare	National Institute for Health and Care Excellence	Public	High
Operational Design Domains	Engineering	(company-specific)	Public	High
BBC Content Standards	Media and Communications	Office of Communications	Public	High
Consolidated Standards of Reporting Trials (CONSORT)	Healthcare	CONSORT Group	Public	High
IAEA Safety Standards	CBRN	International Atomic Energy Agency	Public	Extreme
Responding To A CBRN Event: Joint Operating Principles for the Emergency Services	CBRN	Joint Emergency Services Interoperability Programme	Public	Extreme

Table 2: We used the **criticality levels** component of C3F in Figure 2 to classify a wide variety of standards across domains and sectors based on their sensitivity which translates into the margin of permissible errors a hypothetical GenAI model can potentially commit when assisting with standard compliance tasks. We include supporting information including the organization or initiative which led to the development of the standard as well as its accessibility (Public or Subscription). For standards classified from the domains of healthcare and engineering, we obtained direct recommended assessments from expert practitioners and professionals with respect to the criticality classification criteria in C3F.