

# Rethinking Standard Compliance in the LLM Era

Joseph Marvin Imperial<sup>Ω,Δ</sup> Matthew Jones<sup>Δ</sup> Harish Tayyar Madabushi<sup>Δ</sup>

<sup>Δ</sup>University of Bath, UK <sup>Ω</sup>National University, Philippines

[jmri20@bath.ac.uk](mailto:jmri20@bath.ac.uk) [prpmdj@bath.ac.uk](mailto:prpmdj@bath.ac.uk) [htm43@bath.ac.uk](mailto:htm43@bath.ac.uk)

## Abstract

We are witnessing an emerging *paradigm shift* in which the adoption of large language models (LLMs) has increased significantly across high-stakes, standards-driven industries, particularly in engineering, healthcare, education, and law. In line with EMNLP 2026's theme of "*New Missions for NLP Research*", this position paper aims to initiate civil discussion on how the use of LLMs is shaping the future of standard compliance and introduces *standard alignment*, a targeted research direction towards making LLMs verifiably comply with domain standards. To support the discussion, we propose the CRITICALITY AND COMPLIANCE CAPABILITIES FRAMEWORK (C3F) to assess the criticality levels of various standards and pair them with the current compliance capabilities of state-of-the-art LLMs as evidenced by the literature. Finally, we argue that advancing standard alignment is key to ensuring safe and trustworthy integration of LLMs across standards-driven industries.

## 1 Introduction

The industries of the modern world rely on systematic processes for the efficient production of goods and services guided through **standards**. According to the International Organization for Standardization (ISO)<sup>1</sup>, standards refer to a general form of document-form specifications, rules, and guidelines specialized across various domains and sectors such as healthcare, education, engineering, law, sciences, security, and defense. For example, in the aviation and aerospace industries, any technical instruction manual produced must conform to recognized constrained language standards such as ASD-STE100 Simplified Technical English (STE)<sup>2</sup> developed and maintained by the Aerospace, Security and Defence Industry Association of Europe

<sup>1</sup><https://www.iso.org/standards.html>

<sup>2</sup><https://www.asd-ste100.org/>

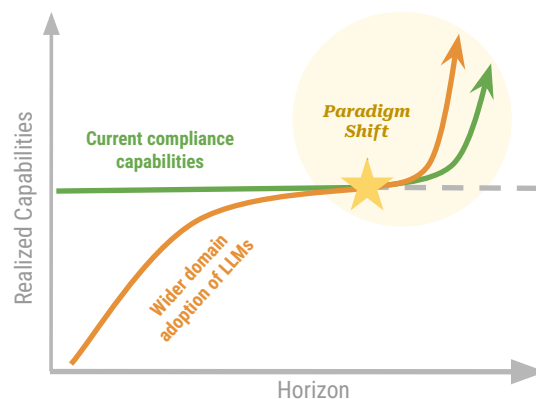


Figure 1: We describe an emerging paradigm shift where LLMs are being explored to assist with compliance tasks. By advancing standard alignment research, we anticipate strengthened compliance capabilities in parallel with the increased safe and trustworthy integration of LLMs across standards-driven industries.

(ASD, formerly AECMA). In the education domain, teachers in charge of developing curriculum materials must follow standards such as the Common European Framework of Reference for Languages (CEFR)<sup>3</sup> in Europe to produce high-quality classroom content.

At its core, standards are typically composed of carefully defined technical *specifications* for measurements and design that can be used to validate regulatory and operational compliance (ASTM International, 2025). In recent years, experts, practitioners, and users across various domains and sectors have shown interest in adopting LLMs due to their capabilities, including the ability to follow complex instructions and generate human-like writing. For example, a survey by the Department for Education revealed that 62% of primary and secondary teachers in the UK have reported using LLM-based tools to help create new educational content and lessons for classroom use (Department

<sup>3</sup><https://www.coe.int/en/web/common-european-framework-reference-languages>

for Education, 2024). Complementary to this, recent studies on controlled text generation show that commercial and open-weight LLMs such as Llama2 (Touvron et al., 2023) and GPT-4 (Achiam et al., 2023) can be systematically steered with prompts to produce high-quality content and obtain high expert human evaluations (Imperial et al., 2024; Malik et al., 2024). Collectively, we see these shifts as a promising research direction, which we refer to as **standard alignment**, in which the goal is to train LLMs to learn to *comply* with standards and their constraints when performing tasks.

In this paper, we analyze a changing landscape—a *paradigm shift*—in regulatory and operational compliance through standards. We present case studies from various sectors that support the paradigm shift, including healthcare, education, safety, finance, and engineering. To further the discussions on how we approach standard alignment, we introduce C3F, a joint high-level CRITICALITY AND COMPLIANCE CAPABILITIES FRAMEWORK for assessing the criticality levels of 30 diverse standards used across various sectors and domains, as well as the current compliance capabilities of 14 general and specialized state-of-the-art LLMs. In summary, we take the following position that *advancing standard alignment research is key to ensuring safe and trustworthy integration of LLMs in standard-driven industries*.

## 2 The Standard Operational Compliance Landscape

To set the stage, we discuss preliminary information covering the common definitions and processes comprising standard development, including the entities involved in their conception and their various purposes.

### 2.1 Defining A Standard

Standards are established to provide specifications in natural language for measuring quality and proof of compliance with regulations. In line with this, we consider standards to be *expert-defined* documents that specify measurable requirements and regulatory orders for compliance by users, industries, and organizations (Demortain, 2008).

To accommodate the diverse and overlapping notions of standards, this paper does not limit the scope of standards to those created by international and regional standard-developing organizations (SDOs), such as ISO, IEEE, ETSI, or NIST.

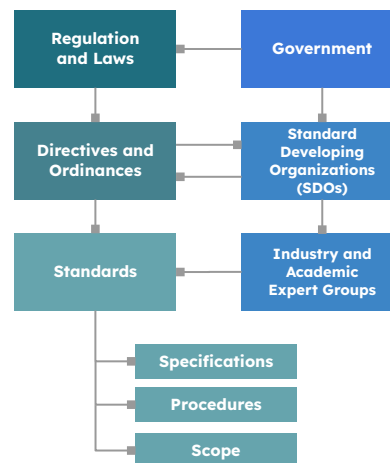


Figure 2: The general process and anatomy of a standard. Standards can be created either by government and regulatory bodies as a product of legislation or through industry associations and academic expert groups to ensure interoperability and quality of systems and processes.

We also include formally documented rules coined under related terms like *frameworks*, *guidelines*, *policies*, *constitutions*, and *checklists*, which are often used in interdisciplinary domains as forms of standards themselves since they observe an 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:

1. **Anatomy:** Standards are composed of well-defined specifications and procedures documenting measurable requirements for a given product or process.
2. **Purpose:** Standards serve specific purposes for various domains and sectors, such as introducing a formal language of communication, ensuring compatibility checks, enforcing a list of rules, and assuring compliance with policies or regulations.
3. **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 government initiatives, regulatory bodies,

unions, organizations, private industries, or expert groups. In legislation, a standard may be paired with specific laws as a **co-regulation tool** that, if successfully met, can serve as *compliance with related state or nationwide jurisdictions* (Pouget and Zuhdi, 2024). These are often called **regulatory standards**, where an example includes the well-known ISO/IEC 27001:2022, which defines measures for Personally Identifiable Information (PII) controllers and processors to implement an information security management system (ISMS) in response to the EU General Data Protection Regulation (GDPR) on data privacy and protection (ISO/IEC 27001:2022). A hierarchical visualization of standard development is shown in Figure 2.

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*. These are often known as **content-based standards** that specify measurable properties of artifacts for consistency. A well-known example is the Web Standards developed by the World Wide Web Consortium (W3C)<sup>4</sup>, which maintains all technical specifications, guidelines, and protocols for web-based technologies, including HTML.

### 3 Paradigm Shift with Large Language Models

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 and practices (Kuhn, 1970; Tapscott and Caston, 1994). In this paper, the emerging paradigm shift we examine centers on how experts, practitioners, and users actively explore ways to integrate standard specifications into LLMs to support compliance-based tasks. We encapsulate this projection in Figure 1.

#### 3.1 Shift in Standard Conformity Assessment

We highlight notable works across various domains in automating standard conformity or compliance assessment with LLMs below:

**Data Privacy Regulations.** Regulations for protecting private and personal information, such as the General Data Protection Regulation (GDPR) (Regulation, 2016) for the EU and the Health Insurance Portability and Accountability Act (HIPAA)

(Act, 1996) for the US, have served as references for automating compliance checks using LLMs. Recent works, such as PRTs (Imperial and Madabushi, 2025), GoldCoin (Fan et al., 2024) and PrivaCI-Bench (Li et al., 2025a), have evaluated a number of LLMs, including Llama2, DeepSeek-R1, and GPT-4, for privacy violation classification, using HIPAA as a reference. Likewise, Zoubi et al. (2024) and Zhu et al. (2024) introduced PrivT5 and LegiLM, new specialized finetuned models trained on compilations of GDPR-related legal content and reported state-of-the-art performance on tasks including agreement validation and clause retrieval.

#### Financial and Accounting Report Standards.

The International Financial Reporting Standards (IFRS) provide a standardized method for evaluating a company’s financial performance, ensuring compliance with national and international regulations (Posner, 2010). The work of Berger et al. (2023) 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-only applications, LLMs have also been explored and have shown promising results for compliance assessment in vision-language settings. Hildebrandt et al. (2024) examined the use of OpenAI’s ChatGPT-4V (OpenAI, 2023) and Vicuna (Chiang et al., 2023) integration to check the compliance of compiled sensor image data from self-driving cars with respect to Operational Design Domains (ODDs). ODDs are documented standards provided by manufacturers (e.g., Tesla or GMC) that describe specific conditions under which a self-driving car may operate safely and within its designed function.

#### 3.2 Shift in Standard-Aligned Generation

We highlight previous studies that focused on improving LLMs’ capabilities to generate content that conforms to the standards automatically:

#### Education and Language Proficiency Frameworks.

Content-based standards serve as a meter to ensure that classroom resources, such as reading and activity books, meet specific research-

<sup>4</sup><https://www.w3.org/standards/>

based quality criteria (La Marca et al., 2000; Sadler, 2017). Content standards such as the Common European Framework of Reference for Languages (CEFR) are among the most widely used resources for automatic educational content generation using LLMs. The combined recent works of Imperial et al. (2024), Malik et al. (2024), and Glandorf and Meurers (2024) 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, structure, and levels of meaning. Experts in language testing using CEFR have also favored how GPT-4 can achieve a certain level of completeness, fluency, and coherence in generated texts (Imperial et al., 2024).

#### **Medical Reporting and Appraisal Standards.**

Checklists and reporting standards drive high-quality documentation and appraisal in the medical literature. Sanmarchi et al. (2024) explored ChatGPT’s capabilities to reformulate the STROBE checklist and analyze epidemiological studies on COVID-19 vaccinations across 68 countries. The results supported ChatGPT’s potential as an assistant in setting up epidemiological observational research, but cautioned against its tendency to produce inconsistent responses when analyzing methods. Likewise, Muluk (2024) used ChatGPT to customize checklists for managing patient-specific musculoskeletal injections that conform to the METRICS standard, thereby supporting the model’s potential to streamline easily verifiable tasks in clinical practice.

**Industry Safety Policies.** Recent work has explored integrating industry safety policies into LLMs to improve their ability to generate content that adheres to these specifications. Examples of these include Anthropic’s Constitutional AI (Bai et al., 2022), OpenAI’s Deliberative Alignment (Guan et al., 2024), and Cos-Align (Zhang et al., 2024). For these methods, LLMs are optimized to interpret the company’s safety specifications and reason over them when responding to potentially harmful prompts. One advantage of OpenAI’s Deliberative Alignment method is that its models have been optimized to identify which policy specifications are relevant to produce a safe and compliant response, rather than reviewing the full policy at every iteration.

## **4 Opportunities with Standard Alignment Research**

We provide an in-depth discussion of the technical advantages that make standard alignment a promising research direction.

### **Standard Alignment Can Improve Oversight, Transparency, and Auditing**

In high-stakes domains, human oversight of AI-based systems or interfaces is crucial for transparency, auditing, and governance (Kolt, 2025; Bowman et al., 2022; Kenton et al., 2024), and is now a major requirement for high-risk systems by recent AI regulations, such as the EU AI Act (European Parliament and Council, 2024). To address this, standard alignment can be viewed as another form of *practical oversight method* where users can control how LLMs produce outputs by updating the specifications of standards, and being able to trace back deviations. This feature is extremely valuable in areas such as bias mitigation (Gallegos et al., 2024), fairness evaluations (Teo et al., 2024), and domain-specific tasks related to healthcare, finance, and legal decision-making (Bowman et al., 2022; Meskó and Topol, 2023; Mökander et al., 2023).

### **Standard Alignment Can Strengthen User Trust**

Building user trust is considered one of the most elusive challenges in the design of AI-based systems (Riegelsberger et al., 2005; Kizilcec, 2016; Schmidt et al., 2020). Bridging the same rules that domain experts follow to control LLMs 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 (Chen et al., 2023) to complete or assist with their tasks, simply by knowing that the model has undergone training on massive collections of clinical guidelines and medical papers with which the expert is familiar. This level of transparency, provided to domain experts, can be highly beneficial for earning and strengthening user trust (Kizilcec, 2016).

### **Standard Alignment Can Reduce Risk of LLM Inaccuracies**

The results of the State of AI in 2024 by McKinsey revealed that *inaccuracy*—the tendency of LLMs to produce factually incorrect and unexpected results—is one of the highest risk factors hindering adoption across major organizations and industries. Such risks can cause a domino effect, including losing user trust, causing harm, and financial losses for consumers and businesses

if not properly mitigated and controlled (McKinsey & Company, 2024). Thus, the standard alignment of LLMs 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 (Imperial and Tayyar Madabushi, 2023; Imperial et al., 2024), legal (Zhu et al., 2024; Fan et al., 2024), and medicine (Chen et al., 2023).

## 5 CRITICALITY AND COMPLIANCE CAPABILITIES FRAMEWORK (C3F)

The level of compliance requirements varies by standard and depends on its purpose and scope. 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 experts, practitioners, and users exploring the use of LLMs as assistants for compliance-based tasks, it is crucial to understand both the documented capabilities of the LLMs and the level of criticality in standard compliance they use as a benchmark for success or failure.

To address this challenge, we propose C3F, a joint high-level CRITICALITY AND COMPLIANCE CAPABILITIES FRAMEWORK in Figure 3, which aims to provide guidance in categorizing the current compliance capabilities of modern LLMs in standard alignment as well as assess the different levels of criticality of standards across domains and sectors.

### 5.1 Assessment of Compliance Capabilities

We define **compliance capabilities** as an *aggregation of a model’s documented capabilities for compliance-based tasks* across various publications and recognition from the NLP and interdisciplinary communities. These capabilities can range from producing standard-compliant outputs based on simple to complex instruction-following to generalizing domain knowledge and producing reasoning steps to justify decisions. For C3F, we propose a high-level, four-stage assessment scheme to measure an LLM’s compliance capabilities in an increasing linear fashion, as shown in Figure 3. We draw motivation on this particular skill assessment of LLMs from previous works such as Morris et al. (2024) for levels of general intelligence and Eiras et al. (2024) for risks and openness.

We anchor each level on documented evidence

from the literature. For the **Baseline**, we consider general off-the-shelf LLMs that can apply explicitly provided specifications from established standards such as coding conventions and checklists (Siddiq et al., 2023; Sanmarchi et al., 2024). LLMs classified as **Specialized** are those that have been domain-adapted through finetuning on standard or regulatory data, or through tool augmentation, allowing them to outperform their base variants on the domain’s compliance tasks, such as clinical guide conformity (Chen et al., 2023) and privacy violation detection (Fan et al., 2024; Hu et al., 2025). For **Advanced**, we consider LLMs that perform comparably with domain experts and can justify which specifications apply when queried, as documented for frontier LLMs that reason over standard and policy clauses and follow complex domain instructions (Imperial and Madabushi, 2025, 2026). Lastly, we consider **Adaptive** as the final level an LLM can obtain, which should generalize compliance across multiple domains and adapt to evolving or previously unseen standards without retraining. No existing LLM is currently classified at this level, but we nonetheless retain this level to situate future progress rather than to assert an empirical threshold (Morris et al., 2024).

### 5.2 Assessment of Criticality Levels

We define the **criticality levels** of standards as *a measure of their task-based sensitivity*, which can be determined by the allowable margin of error for a hypothetical LLM assisting with standard compliance tasks. We also propose a similar high-level four-stage assessment scheme illustrating the reduction of allowable errors as the criticality of standards increases, as shown in Figure 3.

In assessing the criticality of standards, we should consider two variables: the *consequences of harm* and the *scale of harm*. The former describes the level of potential damage that non-compliance with standards can cause, while the latter considers the number of people who might be harmed by an error resulting from non-compliance (Al Atassi et al., 2024). For example, an error in a patient-facing scenario in healthcare, due to the misinterpretation of a standard guideline, might endanger one person. In contrast, an error in a nuclear energy scenario might endanger or kill thousands. Both can potentially lead to death, but the scale varies between the two.

In considering these two variables with Figure 3, **Minimal** standards are the least sensitive,

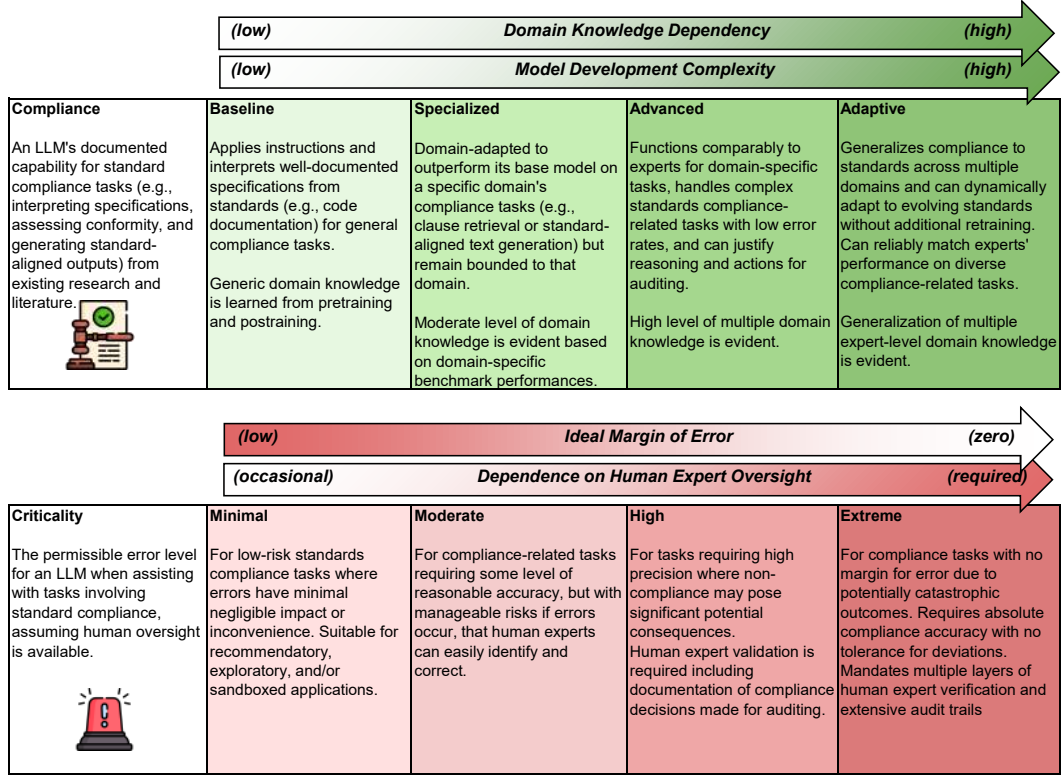


Figure 3: **The CRITICALITY AND COMPLIANCE CAPABILITIES FRAMEWORK (C3F)**. We introduce a joint high-level framework as a guide for assessing current state-of-the-art general and specialized LLMs based on their **(Top)** documented **compliance capabilities** in assessing compliance with standards from previous literature, as well as **(Bottom)** the estimated **criticality levels** of standards from domains and sectors based on the permissible error level an LLM can commit and the potential consequences in the case of non-compliance.

where non-compliance has low-impact risks and can be detected with rule-based checkers. For **Moderate**, there may be some manageable risks with non-compliance that domain experts can readily identify and correct. Standards classified as **High** now require high accuracy and expert validation, with properly documented decisions, since non-compliance may pose significant consequences. Examples in this category include compliance with data privacy and protection laws such as GDPR and HIPAA. Lastly, the highest criticality level a standard can be assigned is **Extreme**, reserved for standards where no margin of error is allowed and non-compliance may result in catastrophic, potentially irreversible consequences. This includes standards under the chemical, biological, radiological, and nuclear (CBRN) umbrella and the IAEA safety standards, an area already drawing preliminary LLM research (de Costa et al., 2024; Hirata et al., 2024).

### 5.3 Intended Users and Application

Our intention in developing C3F is to treat it as a *pre-deployment decision framework* for domain experts, practitioners, researchers, and regulated entities who are interested in using LLMs to automate tasks that require compliance with standards based on their criticality and sensitivity. C3F aims to help users decide which candidate LLMs they have as options, how performant they are on specific tasks, and the extent of human oversight they require, based on documented findings from the existing literature. The decision support judgment from using C3F can span throughout the whole LLM integration stack, including model development, evaluation, pre- and post-deployment, and governance as needed.

## 6 Case Studies

We demonstrate the application of C3F to assess LLMs' compliance capabilities and the criticality of standards. We note that these are high-level, illustrative judgments of how C3F can be applied

| MODEL FAMILY                       | DOMAIN            | OPENNESS    | COMPLIANCE  | EVIDENCE                                                                                                 |
|------------------------------------|-------------------|-------------|-------------|----------------------------------------------------------------------------------------------------------|
| GEMINI SERIES (3, 2.5)             | General           | Commercial  | Advanced    | Cui et al. (2025); Wang et al. (2026); Marino et al. (2025)                                              |
| O-SERIES (O3, O4-MINI)             | General           | Commercial  | Advanced    | Guan et al. (2024); Li et al. (2025b)                                                                    |
| GPT SERIES (5, 4o)                 | General           | Commercial  | Advanced    | Albuquerque and Gomes Dos Santos (2024); Roberts et al. (2023); Zeng et al. (2025); Wang and Chen (2026) |
| DEEPSEEK SERIES (V3.1, R1)         | General           | Open Weight | Advanced    | Guo et al. (2025); Li et al. (2025a); Wang et al. (2026)                                                 |
| CLAUDE SERIES (OPUS 5, SONNET 4.5) | General           | Commercial  | Advanced    | Bai et al. (2022); Zeng et al. (2025); Marino et al. (2025)                                              |
| MEDITRON 70B                       | Healthcare        | Open Code   | Specialized | Chen et al. (2023); Fast et al. (2024); Li et al. (2025b)                                                |
| GOLDCOIN-LLAMA                     | Healthcare, Legal | Open Code   | Specialized | Fan et al. (2024)                                                                                        |
| SAULLM 54B                         | Legal             | Open Code   | Specialized | Colombo et al. (2024); Marino et al. (2025)                                                              |
| STANDARDIZE-LLAMA                  | Education         | Open Code   | Specialized | Imperial et al. (2024)                                                                                   |
| FOUNDATION-SEC-8B                  | Cybersecurity     | Open Weight | Specialized | Kassianik et al. (2025)                                                                                  |
| LLAMA SERIES (4, 3.3)              | General           | Open Weight | Baseline    | Touvron et al. (2023); Berger et al. (2023); Wen et al. (2024)                                           |
| GEMMA SERIES (3, 2)                | General           | Open Weight | Baseline    | Fan et al. (2026); Cui et al. (2025); Fast et al. (2024)                                                 |
| MISTRAL SERIES (3, LARGE)          | General           | Commercial  | Baseline    | Jiang et al. (2023); Fan et al. (2024); Zeng et al. (2025)                                               |
| QWEN SERIES (3, 2.5)               | General           | Open Weight | Baseline    | Li et al. (2025a); Fan et al. (2026); Wen et al. (2024)                                                  |

Table 1: We used the **compliance capabilities** component of C3F in Figure 3 to assess the 14 models and model families. We also include information on the domains they are specialized in, accessibility, and supporting evidence of their respective compliance capabilities from the literature. Note that the models and their current compliance capability labels in this table serve as our illustrative, non-exhaustive judgments and can be extended or reassessed as supporting literature is released.

to assess LLMs and standards, based on evidence from existing work, but should not be taken as absolute, validated measurements of compliance.

### 6.1 Non-Specialist Assessment of Globally Recognized LLMs and Standards

We assess 14 globally recognized frontier and specialized LLMs. We classify all top-of-the-line frontier LLM families (e.g., Gemini, GPT, Claude, DeepSeek) in the **Advanced** category due to their high performance on general and domain-specific LLM benchmarks, which often translates to substantially strong performance in complex compliance-related tasks supported as demonstrated in literature (Hui et al., 2026; Marino et al., 2025; Zeng et al., 2025; Yao et al., 2025; Imperial and Madabushi, 2025). LLMs under the **Specialized** category are often developed and re-

leased by academic stakeholders or in collaboration with sector-specific industries. These LLMs often exhibit significantly improved performance across domain-specific tasks due to additional fine-tuning or optimization with domain-specific data (e.g., massive legal data used by Colombo et al. (2024) for instruction-tuning SaulLM-54B) using LLMs from the **Baseline** category. We report the long table of annotated models in Table 1.

On the standards side, we used the standard criticality level component of C3F to evaluate 15 non-specialist standards, guidelines, and frameworks across coding, media, and software technology. We consider these standards to be non-specialist in this paper’s context, given that they are generally known to the public and frequently applied in low-stakes scenarios (e.g., developing a website that meets Section 508 accessibility stan-

dards) without major serious consequences for non-compliance. Our assessments for these standards are based solely on the tasks for which they are frequently used, such as content creation, documentation enhancements, and compatibility checks, and do not apply to specific edge cases. We report the long table of annotated standards in Table 2.

## 6.2 Specialist Assessment of Healthcare and Pharmaceutical Standards

We asked one of the authors of this work, a pharmacist and researcher in medication safety and clinical guidelines design in the Department of Life Sciences at the University of Bath, UK, to provide a preliminary expert-level assessment of the criticality of the standards used and recognized in the healthcare and pharmaceutical industries. The expert agreed with the purpose of the two criticality variables mentioned previously in Section 5 which are consequences of harm and the scale of harm, and considered this in their evaluation.

We report the 15 specialist high-stakes standards classified by the expert in Table 3 in the Appendix. From the result, we observe that the majority of healthcare-related standards and guidelines fall into the **High** category, emphasizing the significant criticality and high level of oversight required when using them while two were only classified as **Moderate**, as they are only checklist-based standards used in quality assessment. Aside from the evaluation, we also requested that the expert provide verbatim feedback on the practical utility of C3F. We highlight the actual feedback below:

“Currently, there is a lot of interest and unstandardised research in the processing and provision of various types of healthcare information with AI. As this is not an area of expertise for most healthcare professionals, a framework such as C3F is *potentially very useful*. It allows healthcare professionals to link their **assessment of criticality** to AI experts’ **assessments of compliance**. This will support the selection of an appropriate model and workflow for each application and help to standardise the development of this new field.”

— Pharmacist and Medication Safety Researcher

Overall, we view this assessment and feedback as strong support for our position and the usability of C3F in setting a precedent for discussing how LLMs can be applied in standards-driven domains, particularly in the healthcare and pharmaceutical industries and beyond.

## 7 Risks, Responsibilities, and Recommendations

We outline the potential risks and key responsibilities of each stakeholder group and provide recommendations that align with evolving practices in regulatory and operational 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 (Aghion et al., 2023), and the *risk of regulatory gaps* due to vague stipulations and the unrealistic technical feasibility of legislation (Pouget, 2023). To address these, we recommend periodic *regulatory and policy adaptation* to revisit existing regulations and critically discuss how the evolving capabilities and roles of LLMs impact 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 significant job augmentations with the use of LLMs (e.g., physicians now using LLMs to support clinical decision-making), to uphold ethical principles and ensure human oversight (Hashem et al., 2024).

**For Standard-Developing Orgs, Industry, and Academic Expert Groups** Rethinking the conventional standard development process (see Figure 2) due to the shift in compliance practices with LLMs 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 LLMs. We recommend establishing a dedicated committee responsible for updating and initiating maintenance revisions of existing published standards. These domain-specific committees may conduct their own studies and collaborate with other stakeholders to *closely monitor current practices, available tools, and acknowledge limitations* in using LLMs for standard compliance (Manheim et al., 2024). Similarly, to boost advancements in LLMs research with standard compliance, we rec-

ommend open-sourcing machine-readable format of standards along with producing gold-standard compliant exemplars for the research community.

### For LLM Researchers and Model Developers

As more LLMs are used in critical high-risk domains such as healthcare or law, model explainability and interpretability will be an important research direction. Increased scientific works in these areas are crucial to reduce 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 LLMs. Since standards can be considered as co-regulation tools, developing novel approaches such as automatic audit trail generation for decisions made by standard-aligned LLMs and providing human-readable explanations is vital for their wider interdisciplinary and multi-sectoral adoption (Kolt, 2025; Mökander et al., 2023; Song and Shmatikov, 2019).

### For Regulated Entities, Practitioners, and Users

The final stakeholder group that will arguably experience the most significant impact from LLMs is the regulated entities, experts, practitioners, and users. Our most important recommendation is to *practice and uphold the highest form of responsibility and accountability* in using LLMs to comply with regulatory and operational requirements in standards (Coeckelbergh, 2020). To reduce the *risk of misuse and over-reliance*, we also recommend that users build a *solid, domain-specific foundation of AI literacy*, which is expected from an ideal domain expert who acknowledges both advantages and disadvantages of available technical resources used in their domain.

## 8 Alternative Views

While we emphasized and supported our position in the previous sections, we present two alternative views to foster a healthy discussion of the ideas in this paper.

### Standard Alignment Is Just Safety Alignment.

An alternative view to our position is that standard alignment is a subset of the general *safety alignment* research. While this is a reasonable framing, conceptually, we consider safety alignment a *special case* of standard alignment, focusing on aligning models to develop-created AI constitutions and specifications. Standard alignment, on the other

hand, *generalizes to externally authored, version-controlled, and domain-specific specifications* such as CEFR for education, IFRS for accounting, and STROBE for healthcare. We emphasize that treating standards as the broader category is more appropriate for the research direction we advocate. Likewise, standard alignment can be viewed as part of the effort to align AI systems with legal and normative frameworks (Kolt et al., 2026; O’Keefe et al., 2025). Likewise, collapsing these multi-sectoral standards into a single “safety” problem risks dissolving domain-specific context and erasing the role of external domain experts who author and maintain these standards.

### High-Stakes Standard Compliance Should Be Left with Experts.

Another alternative view to our position is that any compliance-related tasks using high-stakes standards (e.g., IAEA standards in nuclear safety) should be left completely to human experts. We strongly agree with the notion that human experts must remain central in these domains, as reflected in C3F, which mandates expert validation and audit trails. Our position, however, aims to use LLMs as first-line support for repetitive, verifiable tasks under human oversight, rather than as autonomous decision-makers. Moreover, as discussed in Section 3, practitioners, even in high-stakes domains, recognize the utility of LLMs. Hence, *the question should not be whether LLMs will be used at all, but whether their use will be guided by thorough standard alignment research.*

## 9 Outlook

LLMs are 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 this transformation as a paradigm shift and discuss how supporting it can be a fundamental step toward *controlled and safe multi-sectoral adoption of LLMs by advancing standard alignment research*. Likewise, we make **two calls to action** across all stakeholder groups involved in the development of standards and LLMs:

1. We encourage all stakeholders to collectively recognize the paradigm shift introduced by LLMs, which can serve as a powerful tool for assisting with compliance with regulations, provided they are used responsibly and ethically.
2. We advocate for a global, unified effort among stakeholders to ensure that LLMs remain a force for good in regulated domains.

## Limitations

With respect to shaping the position of the paper, our proposed C3F is intended to be a *high-level* pre-deployment decision framework to initiate discussions on preliminary assessment of characteristics of capabilities of LLMs and criticality of standards for all concerned stakeholders, including regulatory bodies, industry and academic standard developers, researchers, practitioners, and users. Thus, any form of quantitative validations is purposefully not covered in this work, as this requires upfront knowledge on what specific standard compliance or standard-aligned generation tasks with respect to a given use case context which is extremely subjective across domains. The framework is not prescriptive and is meant to set the tone of the discussion with this position paper, which will effectively inform and guide domain experts towards important aspects to consider (e.g., criticality, current capabilities, margin of error, compute resources) for compliance-related tasks that leverage LLMs (Homewood et al., 2025).

While we provided a general assessment of model capabilities, cross-checked in the literature, and a specialist’s assessment of the criticality of standards in healthcare, we do not claim any form of inter-rater reliability. We believe establishing reproducible criteria and rubrics is an operational follow-up that maturing standard alignment research will produce in the future. All assessments shown in this work are purely illustrative, non-exhaustive, and may change or be revised as LLM deployment across sectors, domains, and industries progresses.

## Acknowledgments

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, Howard Benn, Julian Padget, James Davenport, Tory Frame, Matthew Hewitt, Marina De Vos, George Fletcher, and Jinha Yoon. We also thank Alastair Hunter and Louise Sanders from UKAS, and Christos Christodouloupoulos from ICO, for the insightful discussions on how compliance assessments work in UK government and industry. 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. This work was also partly

supported by the UKRI Strength in Places Fund MyWorld Project (SIPF00006/1).

## References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. [GPT-4 Technical Report](#). *arXiv preprint arXiv:2303.08774*.
- Act. 1996. [Health Insurance Portability and Accountability Act of 1996](#). *Public law*, 104:191.
- Philippe Aghion, Antonin Bergeaud, and John Van Reenen. 2023. [The impact of regulation on innovation](#). *American Economic Review*, 113(11):2894–2936.
- Rached Al Atassi, Wing Shun Chan, Zivile Jurjonaite, Rahel Kahsay, Emily Samson, and Matthew D Jones. 2024. [Validity and reliability of a method to estimate the potential harm of medication errors by considering both the likelihood and degree of harm](#). *International Journal of Pharmacy Practice*, page riae074.
- Fábio Albuquerque and Paula Gomes Dos Santos. 2024. [Exploring ChatGPT’s capabilities in solving accounting standards problems: the case of IAS 37](#). *Cogent Education*, 11(1):2412492.
- ASTM International. 2025. [Form and Style for ASTM Standards](#). Accessed: 2025-01-23.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. [Constitutional AI: Harmlessness from AI Feedback](#). *arXiv preprint arXiv:2212.08073*.
- Armin Berger, Lars Hillebrand, David Leonhard, Tobias Deuser, Thiago Bell Felix De Oliveira, Tim Dilmaghani, Mohamed Khaled, Bernd Kliem, Rudiger Loitz, Christian Bauckhage, and 1 others. 2023. [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.
- Samuel R Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiušė, Amanda Askell, Andy Jones, Anna Chen, and 1 others. 2022. [Measuring Progress on Scalable Oversight for Large Language Models](#). *arXiv preprint arXiv:2211.03540*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. [Language Models are Few-Shot Learners](#). *Advances in Neural Information Processing Systems*, 33:1877–1901.

- Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, and 1 others. 2023. [MEDITRON-70B: Scaling Medical Pretraining for Large Language Models](#). *arXiv preprint arXiv:2311.16079*.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, and 1 others. 2023. [Vicuna: An open-source chatbot impressing gpt-4 with 90%\\* chatgpt quality](#). See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2(3):6.
- Mark Coeckelbergh. 2020. [Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability](#). *Science and Engineering Ethics*, 26(4):2051–2068.
- Pierre Colombo, Telmo Pires, Malik Boudiaf, Rui Filipe Coimbra Pereira de Melo, Gabriel Hautreux, Etienne Malaboeuf, Johanne Charpentier, Dominic Culver, and Michael Desa. 2024. [SaulLM-54B & SaulLM-141B: Scaling Up Domain Adaptation for the Legal Domain](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Antonia Creswell and Murray Shanahan. 2022. [Faithful Reasoning Using Large Language Models](#). *arXiv preprint arXiv:2208.14271*.
- Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. 2025. [OR-bench: An over-refusal benchmark for large language models](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 11515–11542. PMLR.
- Iria da Cunha. 2022. [Un redactor asistido para adaptar textos administrativos a lenguaje claro](#). *Procesamiento del Lenguaje Natural*, 69:39–49.
- Mishca de Costa, Muhammad Anwar, Daniel Lau, and Issam Hammad. 2024. [Classification of Safety Events at Nuclear Sites using Large Language Models](#). *arXiv preprint arXiv:2409.00091*.
- David Demortain. 2008. [Standardising through concepts: The power of scientific experts in international standard-setting](#). *Science and Public Policy*, 35(6):391–402.
- Department for Education. 2024. [Generative AI in education: educator and expert views](#). Government report, Department for Education.
- Francisco Eiras, Aleksandar Petrov, Bertie Vidgen, Christian Schroeder De Witt, Fabio Pizzati, Katherine Elkins, Supratik Mukhopadhyay, Adel Bibi, Botos Csaba, Fabro Steibel, Fazl Barez, Genevieve Smith, Gianluca Guadagni, Jon Chun, Jordi Cabot, Joseph Marvin Imperial, Juan A. Nolasco-Flores, Lori Landay, Matthew Thomas Jackson, and 5 others. 2024. [Position: Near to Mid-term Risks and Opportunities of Open-Source Generative AI](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 12348–12370. PMLR.
- Sabina Elkins, Ekaterina Kochmar, Jackie CK Cheung, and Iulian Serban. 2024. [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.
- European Parliament and Council. 2024. [Regulation \(EU\) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on Artificial Intelligence \(Artificial Intelligence Act\) and amending certain Union Legislative Acts](#). <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>. OJ L 2024/1689, 12.7.2024, p. 1–88.
- Wei Fan, Haoran Li, Zheyang Deng, Weiqi Wang, and Yangqiu Song. 2024. [GoldCoin: Grounding large language models in privacy laws via contextual integrity theory](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3321–3343, Miami, Florida, USA. Association for Computational Linguistics.
- Yu Fan, Jingwei Ni, Jakob Merane, Yang Tian, Yoan Hermstrüwer, Yinya Huang, Mubashara Akhtar, Etienne Salimbeni, Florian Geering, Oliver Dreyer, and 1 others. 2026. EXam: Benchmarking Legal Reasoning on 340 Law Exams. In *International Conference on Learning Representations*, volume 2026, pages 43451–43489.
- Dennis Fast, Lisa C Adams, Felix Busch, Conor Fallon, Marc Huppertz, Robert Siepmann, Philipp Prucker, Nadine Bayerl, Daniel Truhn, Marcus Makowski, and 1 others. 2024. Autonomous medical evaluation for guideline adherence of large language models. *NPJ Digital Medicine*, 7(1):358.
- Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. [Bias and Fairness in Large Language Models: A Survey](#). *Computational Linguistics*, 50(3):1097–1179.
- Dominik Glandorf and Detmar Meurers. 2024. [Towards Fine-Grained Pedagogical Control over English Grammar Complexity in Educational Text Generation](#). In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 299–308, Mexico City, Mexico. Association for Computational Linguistics.
- Melody Y Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Heylar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, and 1 others. 2024. [Deliberative Alignment: Reasoning Enables Safer Language Models](#). *arXiv preprint arXiv:2412.16339*.

- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z.F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 81 others. 2025. [DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning](#). *arXiv preprint arXiv:2501.12948*.
- Shibo Hao, Tianyang Liu, Zhen Wang, and Zhiting Hu. 2023. [ToolkenGPT: Augmenting Frozen Language Models with Massive Tools via Tool Embeddings](#). *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 36:45870–45894.
- Yasmin Hashem, Saba Esnaashari, David Morgan, James Francis, Alexander Poletaev, Francesca Enock, and Jonathan Bright. 2024. [One in Four UK Doctors Are Using Artificial Intelligence: Exploring Doctors’ Perspectives on AI After the Emergence of Large Language Models](#).
- Julio Hernandez, Delaram Golpayegani, and Dave Lewis. 2024. [An Open Knowledge Graph-Based Approach for Mapping Concepts and Requirements between the EU AI Act and International Standards](#). *arXiv preprint arXiv:2408.11925*.
- Carl Hildebrandt, Trey Woodlief, and Sebastian Elbaum. 2024. [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.
- Kenji Hirata, Yusuke Matsui, Akira Yamada, Tomoyuki Fujioka, Masahiro Yanagawa, Takeshi Nakaura, Rintaro Ito, Daiju Ueda, Shohei Fujita, Fuminari Tatsugami, and 1 others. 2024. [Generative AI and large language models in nuclear medicine: current status and future prospects](#). *Annals of Nuclear Medicine*, pages 1–12.
- Aidan Homewood, Sophie Williams, Noemi Drekler, John Lidiard, Malcolm Murray, Lennart Heim, Marta Ziosi, Seán Ó hÉigeartaigh, Michael Chen, Kevin Wei, Christoph Winter, Miles Brundage, Ben Garfinkel, and Jonas Schuett. 2025. [Third-party compliance reviews for frontier AI safety frameworks](#). *arXiv preprint arXiv:2505.01643*.
- Wenbin Hu, Huihao Jing, Haochen Shi, Haoran Li, and Yangqiu Song. 2025. [Safety compliance: Rethinking llm safety reasoning through the lens of compliance](#). *arXiv preprint arXiv:2509.22250*.
- Jie Huang and Kevin Chen-Chuan Chang. 2023. [Towards Reasoning in Large Language Models: A Survey](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.
- Zheng Hui, Doni Bloomfield, and Noam Kolt. 2026. [Copyright-Bench: Agentic Evaluation of Copyright Law Compliance](#). In *Forty-third International Conference on Machine Learning*.
- Joseph Marvin Imperial, Gail Forey, and Harish Tayyar Madabushi. 2024. [Standardize: Aligning language models with expert-defined standards for content generation](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1573–1594, Miami, Florida, USA. Association for Computational Linguistics.
- Joseph Marvin Imperial and Harish Tayyar Madabushi. 2025. [Scaling Policy Compliance Assessment in Language Models with Policy Reasoning Traces](#). *arXiv preprint arXiv:2509.23291*.
- Joseph Marvin Imperial and Harish Tayyar Madabushi. 2026. [Safer Policy Compliance with Dynamic Episodic Fallback](#). *arXiv preprint arXiv:2601.23094*.
- Joseph Marvin Imperial and Harish Tayyar Madabushi. 2023. [Flesch or fumble? evaluating readability standard alignment of instruction-tuned language models](#). In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 205–223, Singapore. Association for Computational Linguistics.
- ISO/IEC 27001:2022. 2022. Information security, cybersecurity and privacy protection — information security management systems — requirements.
- Hamish Ivison, Yizhong Wang, Jiacheng Liu, Zeqiu Wu, Valentina Pyatkin, Nathan Lambert, Noah A Smith, Yejin Choi, and Hannaneh Hajishirzi. 2024. [Unpacking DPO and PPO: Disentangling Best Practices for Learning from Preference Feedback](#). *arXiv preprint arXiv:2406.09279*.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. [Mistral 7B](#). *arXiv preprint arXiv:2310.06825*.
- Sebastian Joseph, Lily Chen, Jan Trienes, Hannah Göke, Monika Coers, Wei Xu, Byron Wallace, and Junyi Jessy Li. 2024. [FactPICO: Factuality Evaluation for Plain Language Summarization of Medical Evidence](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8437–8464, Bangkok, Thailand. Association for Computational Linguistics.
- Paul Kassianik, Baturay Saglam, Alexander Chen, Blaine Nelson, Anu Vellore, Massimo Aufiero, Fraser Burch, Dhruv Kedia, Avi Zohary, Sajana Weerawardhena, and 1 others. 2025. [Llama-3.1-FoundationAI-SecurityLLM-Base-8B Technical Report](#). *arXiv preprint arXiv:2504.21039*.
- Zachary Kenton, Noah Yamamoto Siegel, Janos Kramar, Jonah Brown-Cohen, Samuel Albanie, Jannis Bulian, Rishabh Agarwal, David Lindner, Yunhao

- Tang, Noah Goodman, and 1 others. 2024. [On scalable oversight with weak LLMs judging strong LLMs](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- René F Kizilcec. 2016. [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.
- Noam Kolt. 2025. Governing ai agents. *arXiv preprint arXiv:2501.07913*.
- Noam Kolt, Nicholas Caputo, Jack Boeglin, Cullen O’Keefe, Rishi Bommasani, Stephen Casper, Mariano-Florentino Cuéllar, Noah Feldman, Iason Gabriel, Gillian K Hadfield, and 1 others. 2026. [Legal Alignment for Safe and Ethical AI](#). *arXiv preprint arXiv:2601.04175*.
- Thomas Kuhn. 1970. The Nature of Scientific Revolutions. *Chicago: University of Chicago*, 197(0).
- Paul M La Marca, Doris Redfield, and Phoebe C Winter. 2000. [State Standards and State Assessment Systems: A Guide to Alignment](#). *Series on Standards and Assessments. Non-Journal*.
- Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and 1 others. 2024. [RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback](#). In *Forty-first International Conference on Machine Learning*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. [Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks](#). *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Haoran Li, Wenbin Hu, Huihao Jing, Yulin Chen, Qi Hu, Sirui Han, Tianshu Chu, Peizhao Hu, and Yangqiu Song. 2025a. [PrivaCI-bench: Evaluating privacy with contextual integrity and legal compliance](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10544–10559, Vienna, Austria. Association for Computational Linguistics.
- Xiaomin Li, Mingye Gao, Yuexing Hao, Taoran Li, Guangya Wan, Zihan Wang, Yijun Wang, and Xupeng Chen. 2025b. [MedGUIDE: Benchmarking clinical decision-making in large language models](#). In *NeurIPS 2025 Workshop on GenAI for Health: Potential, Trust, and Policy Compliance*.
- Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. [Synthetic data generation with large language models for text classification: Potential and limitations](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.
- Zeyi Liao and Huan Sun. 2024. [AmpleGCG: Learning a Universal and Transferable Generative Model of Adversarial Suffixes for Jailbreaking Both Open and Closed LLMs](#). In *First Conference on Language Modeling*.
- Ali Malik, Stephen Mayhew, Christopher Piech, and Klinton Bicknell. 2024. [From tarzan to Tolkien: Controlling the language proficiency level of LLMs for content generation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15670–15693, Bangkok, Thailand. Association for Computational Linguistics.
- David Manheim, Sammy Martin, Mark Bailey, Mikhail Samin, and Ross Greutzmacher. 2024. [The Necessity of AI Audit Standards Boards](#). *arXiv preprint arXiv:2404.13060*.
- Bill Marino, Rosco Hunter, Christoph Schnabl, Zubair Jamali, Marinos Emmanouil Kalpakos, Mudra Kashyap, Isaiah Hinton, Alexa Hanson, Maahum Nazir, Felix Steffek, and 1 others. 2025. [AIReg-Bench: Benchmarking Language Models That Assess AI Regulation Compliance](#). *arXiv preprint arXiv:2510.01474*.
- McKinsey & Company. 2024. [The State of AI in Early 2024: Gen AI Adoption Spikes and Starts to Generate Value](#). Accessed: 2025-01-22.
- Bertalan Meskó and Eric J Topol. 2023. [The imperative for regulatory oversight of large language models \(or generative AI\) in healthcare](#). *NPJ Digital Medicine*, 6(1):120.
- Lester James V Miranda, Yizhong Wang, Yanai Elazar, Sachin Kumar, Valentina Pyatkin, Faeze Brahman, Noah A Smith, Hannaneh Hajishirzi, and Pradeep Dasigi. 2024. [Hybrid Preferences: Learning to Route Instances for Human vs. AI Feedback](#). *arXiv preprint arXiv:2410.19133*.
- Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2022. [Reframing Instructional Prompts to GPTk’s Language](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 589–612, Dublin, Ireland. Association for Computational Linguistics.
- Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2023. [Auditing Large Language Models: A Three-Layered Approach](#). *AI and Ethics*, pages 1–31.
- Tomas Bueno Momcilovic, Dian Balta, Beat Buesser, Giulio Zizzo, and Mark Purcell. 2024. [Developing Assurance Cases for Adversarial Robustness and Regulatory Compliance in LLMs](#). In *2024 IEEE 35th International Symposium on Software Reliability Engineering Workshops (ISSREW)*, pages 209–215. IEEE.

- Meredith Ringel Morris, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. 2024. [Position: Levels of AGI for Operationalizing Progress on the Path to AGI](#). In *Forty-first International Conference on Machine Learning*.
- Selkin Yilmaz Muluk. 2024. [Enhancing Musculoskeletal Injection Safety: Evaluating Checklists Generated by Artificial Intelligence and Revising the Preformed Checklist](#). *Cureus*, 16(5):e59708.
- Cullen O’Keefe, Ketan Ramakrishnan, Janna Tay, and Christoph Winter. 2025. [Law-Following AI: Designing AI Agents to Obey Human Laws](#). *Fordham Law Review*, 94:57.
- OpenAI. 2023. [GPT-4V System Card](#). Accessed: 2025-01-14.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. [Training language models to follow instructions with human feedback](#). *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Elliot Posner. 2010. [Sequence as Explanation: The International Politics of Accounting Standards](#). *Review of International Political Economy*, 17(4):639–664.
- Hadrien Pouget. 2023. [The EU’s AI Act Is Barreling Toward AI Standards That Do Not Exist](#). *Lawfare*. Accessed: 2025-01-24.
- Hadrien Pouget and Ranj Zuhdi. 2024. [AI and Product Safety Standards under the EU AI Act](#). Accessed: 2025-01-07.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. [Direct Preference Optimization: Your Language Model is Secretly a Reward Model](#). *Advances in Neural Information Processing Systems*, 36.
- Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. [In-Context Retrieval-Augmented Language Models](#). *Transactions of the Association for Computational Linguistics*, 11:1316–1331.
- James Reason. 1990. [The contribution of latent human failures to the breakdown of complex systems](#). *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 327(1241):475–484.
- Protection Regulation. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council. *Regulation (EU)*, 679:2016.
- Jens Riegelsberger, M Angela Sasse, and John D McCarthy. 2005. [The Mechanics of Trust: A Framework for Research and Design](#). *International Journal of Human-Computer Studies*, 62(3):381–422.
- Richard HR Roberts, Stephen R Ali, Hayley A Hutchings, Thomas D Dobbs, and Iain S Whitaker. 2023. [Comparative study of ChatGPT and human evaluators on the assessment of medical literature according to recognised reporting standards](#). *BMJ Health & Care Informatics*, 30(1).
- D Royce Sadler. 2017. [Academic achievement standards and quality assurance](#). *Quality in Higher Education*, 23(2):81–99.
- Francesco Sanmarchi, Andrea Bucci, Andrea Giovanni Nuzzolese, Gherardo Carullo, Fabrizio Toscano, Nicola Nante, and Davide Golinelli. 2024. [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.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. [Toolformer: Language Models Can Teach Themselves to Use Tools](#). *Advances in Neural Information Processing Systems*, 36:68539–68551.
- Philipp Schmidt, Felix Biessmann, and Timm Teubner. 2020. [Transparency and trust in artificial intelligence systems](#). *Journal of Decision Systems*, 29(4):260–278.
- Mohammed Latif Siddiq, Beatrice Casey, and Joanna Santos. 2023. [A lightweight framework for high-quality code generation](#). *arXiv preprint arXiv:2307.08220*.
- Congzheng Song and Vitaly Shmatikov. 2019. [Auditing Data Provenance in Text-Generation Models](#). In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 196–206.
- Don Tapscott and Art Caston. 1994. [Paradigm Shift: The New Promise of Information Technology](#). *Economic Development Journal of Canada*, pages 62–66.
- Christopher Teo, Milad Abdollahzadeh, and Ngai-Man Man Cheung. 2024. [On Measuring Fairness in Generative Models](#). *Advances in Neural Information Processing Systems*, 36.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. [Llama 2: Open Foundation and Fine-Tuned Chat Models](#). *arXiv preprint arXiv:2307.09288*.
- Can Wang and Yiqun Chen. 2026. [Evaluating Large Language Models for Evidence-Based Clinical Question Answering](#). *Patterns*, 7(5).
- Jianwei Wang, Mengqi Wang, Yinsi Zhou, Zhenchang Xing, Qing Liu, Xiwei Xu, Binghao Li, Serkan Saydam, Wenjie Zhang, and Liming Zhu. 2026. [LLM-Based HSE Compliance Assessment: Benchmark](#),

- Performance, and Advancements. In *International Conference on Intelligent Systems and Knowledge Engineering*, pages 507–536. Springer.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. [Chain-of-Thought Prompting Elicits Reasoning in Large Language Models](#). *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Bosi Wen, Pei Ke, Xiaotao Gu, Lindong Wu, Hao Huang, Jinfeng Zhou, Wenchuang Li, Binxin Hu, Wendy Gao, Jiaxin Xu, and 1 others. 2024. [Benchmarking Complex Instruction-Following with Multiple Constraints Composition](#). *Advances in Neural Information Processing Systems*, 37:137610–137645.
- Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2025.  [\$\tau\$ -bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains](#). In *The Thirteenth International Conference on Learning Representations*.
- Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. [ZeroGen: Efficient Zero-shot Learning via Dataset Generation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Yi Zeng, Yu Yang, Andy Zhou, Jeffrey Ziwei Tan, Yuheng Tu, Yifan Mai, Kevin Klyman, Minzhou Pan, Ruoxi Jia, Dawn Song, and 1 others. 2025. [AIR-BENCH 2024: A Safety Benchmark based on Regulation and Policies Specified Risk Categories](#). In *The Thirteenth International Conference on Learning Representations*.
- Jingyu Zhang, Ahmed Elgohary, Ahmed Magooda, Daniel Khashabi, and Benjamin Van Durme. 2024. [Controllable Safety Alignment: Inference-Time Adaptation to Diverse Safety Requirements](#). *arXiv preprint arXiv:2410.08968*.
- Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. [LIMA: Less Is More for Alignment](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 55006–55021. Curran Associates, Inc.
- Linkai Zhu, Lu Yang, Chaofan Li, Shanwen Hu, Lu Liu, and Bin Yin. 2024. [LegiLM: A Fine-Tuned Legal Language Model for Data Compliance](#). *arXiv preprint arXiv:2409.13721*.
- Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. [Universal and Transferable Adversarial Attacks on Aligned Language Models](#). *arXiv preprint arXiv:2307.15043*.
- Mohammad Zoubi, Santosh T.y.s.s, Edgar Rosas, and Matthias Grabmair. 2024. [PrivaT5: A generative language model for privacy policies](#). In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pages 159–169, Bangkok, Thailand. Association for Computational Linguistics.

## A Technical Methods Towards Standard Alignment

We provide supplementary information on promising novel technical and algorithmic methods that can be explored (but are not limited to) further improving the standard alignment of LLMs and cater to their inherent complexities as discussed in Section 4. 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 LLMs. 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 (Brown et al., 2020; Mishra et al., 2022). This has been done on works such as the production of high-quality standard-conforming educational content with CEFR and Bloom’s Taxonomy (Imperial et al., 2024; Malik et al., 2024; Elkins et al., 2024) and rewriting complex texts to conform to government-mandated plain languages guidelines (da Cunha, 2022; Joseph et al., 2024). The advantage of ICL is its simplicity, which can easily be explored by non-technical domain users with any arbitrary LLM-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. (2024), 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 (Ouyang et al., 2022) and DPO (Rafailov et al., 2024) combined with supervised finetuning (SFT) to improve the response align-

ment of LLMs 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 (Guan et al., 2024) where an LLM is trained with chain-of-thought (CoT) style prompts (Wei et al., 2022) 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 (Zhou et al., 2023). Nonetheless, combinations of post-training techniques (e.g., SFT + DPO with standard-aligned preference pairs + CoT prompting) are viable approaches for improving an LLM’s standard compliance capabilities.

**Synthetic Data Generation** Synthetic data generation using modern LLMs that have undergone processes such as larger scaling, instruction-tuning, and preference optimization often outperform other data augmentation techniques for downstream tasks (Ye et al., 2022; Li et al., 2023). In standard alignment of LLMs, 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. (2024), 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 (Miranda et al., 2024; Ivison et al., 2024; Lee et al., 2024).

**Retrieval and Tool Augmentation** Augmenting LLMs, 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 (Hao et al., 2023; Schick et al., 2023). 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 an LLM that knows how to call a calculator tool to approximate how long sentences should be generated. Moreover, a search engine tool can also help LLMs access updated versions of standard specifications from its original web sources as prior version checking. On the other hand, another approach that can improve LLMs’ 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 (Lewis et al., 2020; Ram et al., 2023).

**Reasoning Capabilities** Whether LLMs, 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 (Huang and Chang, 2023). Previous research often claims that such capability can be triggered in different ways, such as providing intermediary reasoning steps to prompts (Wei et al., 2022) or using arbitrary models to select and infer reasoning steps from context information (Creswell and Shanahan, 2022). Assuming LLMs can actually reason, in standard alignment, such skill may play an important role in deciding which standard specifications 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 (Guan et al., 2024) 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, LLMs rated Advanced and Adaptive in C3F should document convincing reasoning capabilities across applicable tasks.

**Adversarial and Compliance Robustness Testing** Standard compliance is inherently synergistic and collaborative between regulators and standard developers, as well as regulated entities, practitioners, and users, as this form of ecosystem provides

mutual benefits, including interoperability of processes and conformity to regulations. However, research in improving compliance robustness due to regulatory shifts or changes in constraints and adversarial testing can be extended to standard alignment of LLMs to holistically measure their compliance capabilities under unseen test cases. Very recent works that have led this direction include (Momcilovic et al., 2024), where they proposed a framework leveraging on layered guardrails—inspired by the Swiss cheese model from cybersecurity (Reason, 1990)—that can be integrated with chat-based LLMs to mitigate potential attacks which allows them to maintain compliance to the EU AI Act. On the other hand, (Zou et al., 2023) and (Liao and Sun, 2024) both proposed universal and transferable methods for automatically generating suffixes which can be used for adversarial and robustness of open and closed LLMs. These methods can be seamlessly integrated to form standard-specific adversarial prompt datasets to benchmark compliance capabilities of LLMs as it scales from Baseline to Advanced (and even Adaptive) based on C3F in Section 5.

**Considerations in Future Technical Enhancements** While the scope of this section focuses on evolving approaches in alignment, reasoning, retrieval, data augmentation, tool integration, and benchmarking that can be explored to further enhance the compliance capabilities of LLMs, the central aspect that will remain constant as we progress in the future in this direction is the standards themselves. Standards will always remain a resource solely created by human domain experts that any form of LLMs or technical advancement that will be integrated into LLMs, regardless of any level of measured intelligence, must conform to and remain absolute compliant with if it will be deployed as an assistant to any tasks requiring compliance.

## B Full Assessment Results of LLMs and Standards via C3F

As discussed in Section 5, we provide the full list of the 15 selected foundational and domain-optimized LLMs assessed based on their compliance capabilities and the 30 selected standards from non-specialist and specialist evaluations across multiple disciplines for their criticality levels.

## **C Declaration of LLM Use**

In producing this work, we used Grammarly and Claude Code (Opus 4.8) for minor grammar and spelling corrections, LaTeX table fixes, and visualization assistance. All suggestions from these tools were scrutinized by the authors before being integrated into the paper. No LLM was used in brainstorming, content generation, idea conception, and related literature writing for this work.

| STANDARD                                                    | DOMAIN                   | DEVELOPER                                                              | OPENNESS   | CRITICALITY |
|-------------------------------------------------------------|--------------------------|------------------------------------------------------------------------|------------|-------------|
| Simplified Molecular Input Line Entry System (SMILES)       | Chemistry                | David Weininger                                                        | Public     | Minimal     |
| Associated Press Stylebook                                  | Media and Communications | Associated Press                                                       | Commercial | 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                              | Commercial | Minimal     |
| IFRS Accounting Standards                                   | Finance                  | International Financial Reports Standards                              | Public     | Minimal     |
| FAIR Data Principles Documentation                          | Software and Technology  | GO FAIR Initiative                                                     | Public     | Minimal     |
| 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     |
| Common European Framework of Reference for Languages (CEFR) | Education                | Council of Europe                                                      | Public     | Minimal     |
| Common Core Standards (CCS)                                 | Education                | National Governors Association, Council of Chief State School Officers | Public     | Minimal     |
| ASD-STE100 Simplified Technical English (STE)               | General, Engineering     | Aerospace, Security and Defence Industries Association of Europe       | Public     | Minimal     |
| BBC Content Standards                                       | Media and Communications | Office of Communications                                               | Public     | Moderate    |

Table 2: Our **non-specialist assessment** of the **criticality levels** of 15 globally-recognized low-stakes standards and frameworks across software and technology, media, education, government, and engineering domains based on C3F in Figure 3. We include supporting information, including the organization or initiative that led to the development of the standard, as well as its accessibility (Public or Commercial). Note that this assessment serves as a non-exhaustive illustrative example of how C3F can be used and extended or reassessed as additional standards in the domains are released.

| STANDARD                                                                                 | DOMAIN                              | DEVELOPER                                                              | OPENNESS | CRITICALITY |
|------------------------------------------------------------------------------------------|-------------------------------------|------------------------------------------------------------------------|----------|-------------|
| 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    |
| 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        |
| 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        |
| 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        |
| Consolidated Standards of Reporting Trials (CONSORT)                                     | Healthcare                          | CONSORT Group                                                          | Public   | High        |
| GDPR Documentation Standard                                                              | Legal, Healthcare                   | European Union                                                         | Public   | High        |

Table 3: A healthcare expert’s **specialist assessment** of the **criticality levels** of 15 globally-recognized high-stakes standards and frameworks used in medical and pharmaceutical industries for both professional practices and research based on C3F in Figure 3. Note that this assessment serves as a non-exhaustive illustrative example of how C3F can be used and extended or reassessed as additional standards in the domains are released.