ICS 11.120.10

CCS C10

团体标准

T/SBIAORG 000X—2025

基于大语言模型的药物警戒个例安全性报告人工智能辅助信息提取规范

Specifications for AI-Assisted Data Extracion for Individual Case Safety Reports in Pharmacovigilance Based on Large Language Models

2025-xx-xx发布　　2025-xx-xx实施

上海市生物医药行业协会

目录

前言

本文件按照 GB/T 1.1－2020《标准化工作导则 第 1 部分：标准化文件的结构和起草规则》的规定起草。

本文件由上海市药品和医疗器械不良反应监测中心提出，上海市生物医药行业协会发布。

本文件起草单位：上海市药品和医疗器械不良反应监测中心、上海罗氏制药有限公司、上海商汤科技开发有限公司、上海复宏汉霖生物制药有限公司、上海复星医药（集团）股份有限公司、江苏恒瑞医药股份有限公司

本文件首批执行单位：上海罗氏制药有限公司、上海商汤科技开发有限公司、上海复宏汉霖生物制药有限公司、上海复星医药（集团）股份有限公司、江苏恒瑞医药股份有限公司

本文件主要起草人：许晋、胡骏、巢艾伦、杨依晗、栾国琴、黄天娇、袁博、步伐、浦义虎、韩澎、黎洋本标准为首次发布。

应用大语言模型自动提取个例安全性报告信息

实现结构化数据处理

1．范围

本标准适用于应用大语言模型技术进行药物警戒个例安全性报告数据处理的组织和机构，包括药品上市许可持有人、境外上市许可持有人的境内责任人、药品经营企业等。此外，也可为药品监管部门、医疗机构及相关技术开发商和服务提供商，在开发、实施和使用基于大语言模型的药物警戒数据处理系统或工具时提供参考。

2．规范性引用文件

《药品管理法》

《药物警戒质量管理规范》

《个例药品不良反应收集和报告指导原则》

《个例安全性报告 E2B（R3）区域实施指南》

《生成式人工智能服务管理暂行办法》

《中华人民共和国个人信息保护法》

《中华人民共和国数据安全法》

《中华人民共和国网络安全法》

GB/T 41867—2022 信息技术 人工智能 术语

GB/T 23703.2-2010 知识管理 第 2 部分：术语

GB/T 42135—2022 智能制造 多模态数据融合技术要求

《EMA Guiding principles on the use of large language models in regulatory science and for medicines regulatory activities》

ISO 24611:2012 语言资源管理——形态句法注释框架（MAF）Language resource management — Morpho-syntactic annotation framework (MAF)

3. 术语和定义

下列术语和定义适用于本文件。

3.1 药物警戒 pharmacovigilance

对药品不良反应及其他与用药有关的有害反应进行监测、识别、评估和控制。

［来源：药品管理法］

3.2 大语言模型 Large Language Model,LLM

生成式人工智能的一种，专注于文本生成。

[来源：Guiding principles on the use of large language models in regulatory science and for medicines regulatory activities]

3.3 生成式人工智能技术 generative artificial intelligence

具有文本、图片、音频、视频等内容生成能力的模型及相关技术。

[来源：生成式人工智能服务管理暂行办法]

3.4 微调 fine-tuning

为提升人工智能模型的预测精确度，一种先以大型广泛领域数据集训练，再以小型专门领域数据集继续训练的附加训练技术。

[来源：GB/T 41867—2022]

3.5 多模态数据 multi-modal data

多种形态的数据。

注：包含结构化数据（例如业务系统数据等）、半结构化数据（例如 XML 文件、JSON 文件等）和非结构化数据（例如文本、语音和图像视频等）。

[来源：GB/T 42135—2022]

3.6 知识 knowledge

通过学习、实践或探索所获得的认识、判断或技能。

[来源：GB/T 23703.2-2010]

3.7 知识库 knowledge base

用于知识管理的一种特殊的数据库，以便于有关领域知识的采集、整理以及提取。

3.8 词元 Token

文档中非空的连续字形或音素序列。

[来源：ISO 24611:2012]

3.9 提示信息 prompt

描述生成式人工智能模型应执行任务的文本。提示词包含提供给大语言模型（LLM）的任何文本，但不一定对用户可见。

[来源：Guiding principles on the use of large language models in regulatory science and for medicines regulatory activities]

3.10 提示工程 prompt engineering

人工智能中的一个概念，在提示工程中，执行任务的描述文本会被嵌入到输入中。

4.总体考量

在大语言模型应用系统的开发、部署、验证和质量控制的全生命周期中，应遵循以下原则，以确保安全、负责和有效地使用 LLM 系统。

1) 人工审查和监督 Human Oversight ：确保 LLM 系统在整个开发和使用过程中始终有人工审查和监督。
2) 可问责 Accountability：建立明确的责任分配和归属制度。
3) 透明性 Transparency：保持系统的透明度，并能够向相关方提供准确和及时的信息，包括系统的功能、操作流程、使用情况和局限性等。
4) 可解释性 Explainability：提供 LLM 决策过程的可理解解释，包括完整的日志和记录等，以便在必要时进行审查和分析，确保系统操作的可追溯性。
5) 风险管理：采用基于风险的方法来管理 LLM 应用，将 LLM 系统纳入药物警戒质量管理系统，持续识别、评估并减轻潜在的风险。
6) 数据保护和隐私 Data Protection and Privacy：加强数据保护和隐私保护，确保个人数据和敏感信息的处理符合《中华人民共和国个人信息保护法》、《中华人民共和国数据安全法》、《中华人民共和国网络安全法》以及数据跨境等相关法律法规要求。
7) 治理 Governance：建立有效的治理框架，以监督和规范 LLM 系统的使用。
8) 持续学习 Continuously Learning：鼓励组织和用户持续学习和适应如何合理合规使用 LLM 系统，以最大化 LLM 的使用价值。

5.大语言模型在个例安全性报告信息提取场景下的系统实施

5.1 模型部署和架构设计

5.1.1 大语言模型基模型的选择

1) 宜选择与应用场景匹配的大语言模型。例如，个例报告源文件可能较长，宜优先考虑在长文本能力具有优势的大语言模型，大语言模型支持的上下文 Token 宜不低于 128K；个例报告信息提取的规则较多，应优先选取在逻辑推理方面能力较强的大模型；个例报告源文件包含文本、图片、语音等多种格式，宜考虑在多模态数据处理能力具有优势的大语言模型；宜综合应用场景复杂度、技术成熟度、成本预算等因素选择合适的商业版大语言模型或者开源版大语言模型。
2) 宜选择支持持续预训练微调和微调定制化的大语言模型，以更好的匹配药物警戒应用场景。
3) 宜考虑组织或用户对模型部署的具体需求，选择符合需求的大语言模型。

4） 宜选择具备持续学习和更新能力的大语言模型，以满足未来业务的可持续发展。

5.1.2 模型调优

调优的候选技术手段可包括提示工程，知识库管理，模型微调等，旨在提高大模型输出结果的准确率。

1） 提示工程：采用在提示词中加入少量示例、加入思维链引导模型推理、拆分提取任务（如先提取怀疑用药，再提取怀疑用药的属性信息，包括开始日期，适应症等）、提示语结构优化等基于提示语的调优手段，提升大语言模型对任务的理解，降低大语言模型逻辑推理复杂度，从而提高输出结果的准确率。

2） 知识库管理：对于需要补充参考语料（与任务相关的额外背景信息、领域知识或示例等）进行提取的信息，可将补充语料放入外部知识库中。宜参考补充制药企业遵循的的政策法规要求(如《药物警戒质量管理规范》、《个例药品不良反应收集和报告指导原则》、《个例安全性报告 E2B（R3）区域实施指南》等 ） 和领域知识（如 MedDRA(Medical Dictionary for Regulatory Activities)、CTCAE(Common Terminology Criteria for Adverse Events)等）。

3） 模型微调：模型微调通常依赖于至少千条标注好的训练数据，并需通过自动化评估系统来检验微调成效。在决策是否进行微调或确定微调策略时，宜全面评估数据质量、标注成本和测试集的适用性。

5.1.3 模型部署和架构设计

可依据实际需求选择适宜的部署策略，在架构设计时还应考虑数据保护的合规性，确保符合数据安全、隐私保护、伦理、数据跨境等相关要求。

5.2 模型性能评估

大语言模型在药物警戒个例报告信息抽取任务的性能评估宜采用混淆矩阵中的准确率(Accuracy)、精确率(Precision)、召回率(Recall)指标.

5.2.1 评估指标定义

综合考虑药物警戒实际业务的评测数据情况，宜将混淆矩阵的 TP（True Positive）、FN（False Negative）、FP（False Positive）、TN（True Negative）转换为药物警戒场景下的业务术语：正提、漏提、错提、提错和正拒，术语定义和映射关系如下：

1） 正提：源文件信息存在，且被大语言模型正确提取

2） 漏提：源文件信息存在，但未被大语言模型提取出来

3） 错提：源文件信息不存在，但大语言模型提取出了信息

4） 提错：源文件信息存在，但大语言模型提取出了错误信息

5）正拒：源文件信息不存在，大语言模型没有提取出信息

补充说明：

1） 源文件：药物警戒个例安全性报告的原始信息文件，例如文献、电子邮件、来自市场研究项目或患者支持项目的报告表等

2） 源文件信息：药物警戒个例安全性报告所需填写的字段信息，例如患者信息（姓名、性别、出生日期等）、报告人信息（姓名、职业、所在单位、联系电话、电子邮箱等）、药品信息（批准文、商品名、通用名、剂型、用法用量等）、不良反应信息（不良反应术语、严重性、报告人评价等）

	预测正类	预测负类
实际正类	TP （正提）	FN （漏提）
实际负类	FP （错提或提错）	TN （正拒）

5.2 评估指标计算

准确率（Accuracy）$= \frac{TP+TN}{TP+TN+FP+FN} = \frac{正提+正拒}{正提+正拒+漏提+错提+提错}$

注：准确率是最直观的性能指标，它衡量的是模型预测正确的样本占总样本的比例，旨在全面评估模型的整体性能。

精确率（Precision）$= \frac{TP}{TP+FP} = \frac{正提}{正提+错提+提错}$

注：精确率衡量的是模型预测为正类的样本中，实际为正类的比例。精确率关注的是预测结果的准确性，即减少误报。

召回率（Recall）$= \frac{TP}{TP+FN} = \frac{正提}{正提+漏提}$

注：召回率衡量的是所有实际为正类的样本中，被模型正确预测为正类的比例。召回率关注的是模型捕捉到的正类样本的完整性，即减少漏报。

F1 分数（F1 Score）$= (1+\beta^2)\frac{Precision \times Recall}{(\beta^2 \times Precision) + Recall}$

注：F1 分数是精确率和召回率的调和平均数，它试图在精确率和召回率之间找到一个平衡。F1 分数是一个综合指标，适用于需要同时考虑精确率和召回率的场景。

6. 系统应用和质量控制

6.1 制定大语言模型应用系统使用策略

根据系统输出准确率，法规合规性要求，数据准确度要求等方面，进行综合评价，确定大语言模型使用策略。具体可分为基本依赖大语言模型进行个例安全性报告信息的自动提取，大语言模型与人工相结合但以大语言模型为主进行个例安全性报告信息的提取和大语言模型与人工相结合但以人工为主进行个例安全性报告信息的提取。

6.1.1 使用策略建议

模型输出准确率/全部数据点准确率	系统可用性等级	关键数据点准确率*	策略建议
>98%	非常高	不适用	以大语言模型为主，需要人工控制质量
>90%	高	>98%	以大语言模型为主,部分依赖人工，需要人工控制质量
		<98%	以人工为主，部分依赖大语言模型,需要人工控制质量
≥60%；≤90%	中	不适用	以人工为主,部分使用大语言模型与人工相结合方式，需要人工控制质量
<60%	低	不适用	大语言模型与人工相结合但以人工为主或不考虑大语言模型

*注 1：关键数据点包括并不限于不良事件术语、药品信息、患者信息、报告人严重性判断和相关性判断等。

**注 2：上述建议主要适用于有法规递交要求的安全性信息提取。如果所处理数据，只是为内部使用，不涉及任何法规递交，例如既往数据库的数据迁移，则需要全面评估，决定使用策略。

6.2 标准操作流程（SOP）

基于现有流程体系和大语言模型使用策略，建立相关标准操作流程并及时更新，嵌入现有流程体系，确保流程体系的完善性和准确性，从而确保合规性。

6.3 结果复核

应对大语言模型工具输出的结果进行人工复核。可基于实际应用场景和合规性要求，设计复核方式和流程。

6.4 定期回顾和更新

6.4.1 应定期对大语言模型工具的输出结果进行总结和分析，识别精确率较低的数据元素，并采取有针对性的纠正行动，提高大语言模型的提取准确率。

6.4.2 应及时针对大语言模型工具的知识库进行更新和补充。

6.4.3 根据具体情况决定是否对大语言模型进行升级，如决定升级，应开展充分的性能评估后进行升级。

6.4.3 针对大语言模型工具的重要更新和升级后，必要时再次进行综合评估，更新大语言模型使用策略，并相应更新相关 SOP 和具体操作流程。

6.5 文档记录

6.5.1 大语言模型工具的提取结果文档，不应视作源文件，应存档最原始的源文件文档。

6.5.2 大语言模型工具不宜作为个例安全性报告数据存档工具，相关数据应及时清理，但需要大语言模型使用记录等相关事宜的记录。

6.6 培训

按照药物警戒流程体系，针对使用大语言模型工具的人员进行定期培训，确保相关人员按照药物警戒流程体系和相关法规要求开展相关工作。

6.7 合规性

实时监测大语言模型使用过程中的合规性。对使用过程中产生的偏差需要及时记录和纠正，必要时进行相应 CAPA，确保合规性。

6.8 审计

按照药物警戒相关审计要求，定期对大语言模型使用的相关流程进行审计，确保合规使用。

6.9 业务持续性

按照公司或药物警戒部门业务持续性要求，进行业务持续性评估和必要的演练。确保在大语言模型工具使用出现重大问题时，能够正常开展个例安全性报告数据处理工作。

7. 数据安全和隐私保护

7.1 缓解措施

在规划、设计和运行该应用时应充分考虑数据安全和隐私保护。宜实施有效的缓解措施，以防止： a. 提示词注入；b. 不安全的输出处理；c. 数据中毒；d. 模型拒绝服务；e. 供应链漏洞；f. 不安全的插件设计；g. 过度代理；h. 过度依赖；i. 模型盗窃；j. 漂移。缓解控制措施宜包括但不限于 7.1.1-7.1.3 中的描述。

7.1.1 遵循中国关于数据安全与隐私保护方面的法律法规,特别可以参考人工智能方面的法律法规要求，如《生成式人工智能服务管理暂行办法》.

7.1.2 可参考国际组织对于数据信息安全资质的要求：

ISO/IEC 27001:2022 信息安全管理体系

ISO/IEC 29151:2017 个人身份信息保护管理体系

ISO/IEC 27701:2019 隐私信息安全管理体系

ISO/IEC 27017:2015 云安全管理体系

ISO/IEC 27018:2019 公有云个人信息保护管理体系

ISO/IEC 23894:2023 信息技术 - 人工智能 - 风险管理指南

BS 10012 个人信息安全管理体系

7.1.3 应符合组织内部的关于数据安全和隐私保护方面的合规要求。例如某些组织内部会对依据数据的敏感度进行分类（如公开、内部、机密、高度机密）。对高敏感数据实施严格的控制措施，以防止未经授权的访问和泄露。

7.2 基模型云服务提供商的报告 or 证书要求

Cloud 云服务用户数据保护能力检验报告

Cloud 云服务用户数据保护能力检验证书

AIDC 智算中心基础设施支撑平台-网络信息安全等级保护三级

7.3 条件限制

如使用知识产权、个人数据或个人敏感信息等之前，应对使用进行特定的限制或强制性先决条件。

8、利益相关方沟通

大语言模型应用系统使用方应与监管机构、医疗机构等利益相关方保持沟通，确保其了解 LLM 的使用情况和潜在影响，确保 LLM 的应用满足监管要求，并能够支持监管决策。