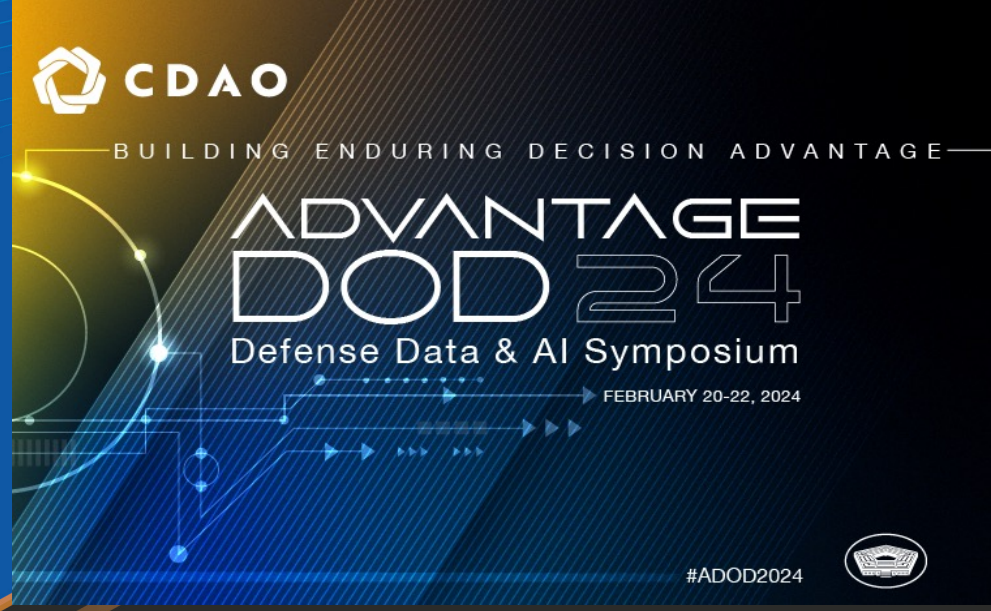
Large Language Model (LLM) Maturity Model

Chief Digital and Artificial Intelligence Office (CDAO)



February 14, 2024

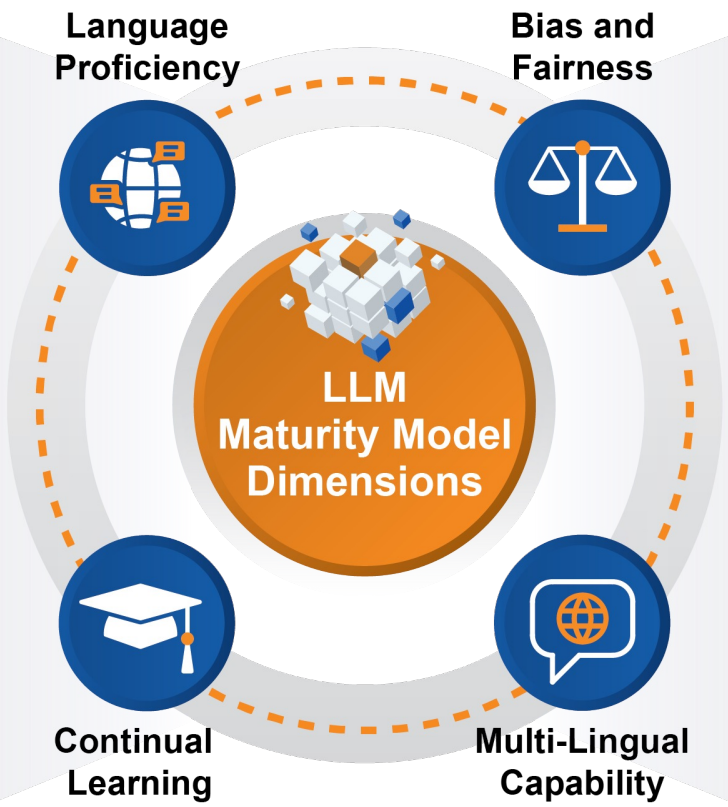
Submitted To
Advantage DoD 2024, CDAO Symposium
osd.pentagon.cdao.mbx.engagement@mail.mil

Prepared By
iWorks Corporation
1889 Preston White Drive, Suite 100, Reston, VA 20191
Phone: (703) 636-6777 | Fax: (866) 574-2310 | www.iworkscorp.com
UEI: ZGHSJN13NVZ5 | DUNS: 16-9988578 | CAGE: 4Q7A1

CMMI DEV/4 Exp. 2023-02-14 / Appraisal #6082 CMMI SVC/3 ISO 9001:2015 ISO 20000:2011 ISO 27001 Certified

LLM MM Level Overview

DOD Use Cases
Conversational chatbot
Summary generation
Question and answer
Analysis
Classification
Insight generation
Language translation
Text Generation
Text Augmentation
Speech Recognition & Synthesis
Image Captioning & Generation
Knowledge Extraction and Representation
Developer-related



LLM Characteristics
Governability
Completeness
Accuracy
Equitability
Consistency
Traceability
Novelty
Fluency
Interactivity
Robustness
Adaptability
Scalability
Efficiency
Security
Privacy
Explainability
Reliability
Usability
Interoperability

Expanded use cases and characteristics for LLM users across the DOD echelons

Extended Use Cases for Wider LLM Usage

Text Generation

- Generation of contextually-relevant text
- Production of high-quality, semantically relevant text

Speech Recog. & Synthesis

- Needed for tasks such as voice assistance, automated dispatch, etc.
- Responding naturally and conversationally

Know. Extraction & Rep.

- Identify relevant information from unstructured data sources, such as text, images, or videos
- Extract and transform as structured knowledge for more efficient and interpretable representation

Text Augmentation

- Generation of new data samples by applying various transformations to existing data samples
- Improves robustness and accuracy

Image Caption & Generation

- Needed to generate human-like descriptions of visual content - descriptive and contextually relevant
- Supports language and visual processing for complex tasks

Developer-Related

- Needed to assess the design, deployment, and use of generative AI technologies responsibly and securely
- Supports mission effectiveness by automating certain tasks, reducing human error, and increasing efficiencies

Extended use cases to support effective LLM usage

LLM Maturity Models - Dimensions

Language Proficiency

- *Accuracy and Fluency*
- *Semantic Understanding*
- *Multilingual Competence*
- *Contextual Appropriateness*
- *User-Friendly Interfaces*
- *Fine-Tuning*

Bias and Fairness

- *Bias Detection and Mitigation*
- *Ethical Guidelines*
- *Data Representativeness*
- *Continuous Monitoring*
- *User Feedback Mechanisms*
- *Interpretability and Transparency*

LLM Maturity Model Dimensions

LLM MM SCORING RUBRIC FOR 4 DIMENSIONS

Continual Learning

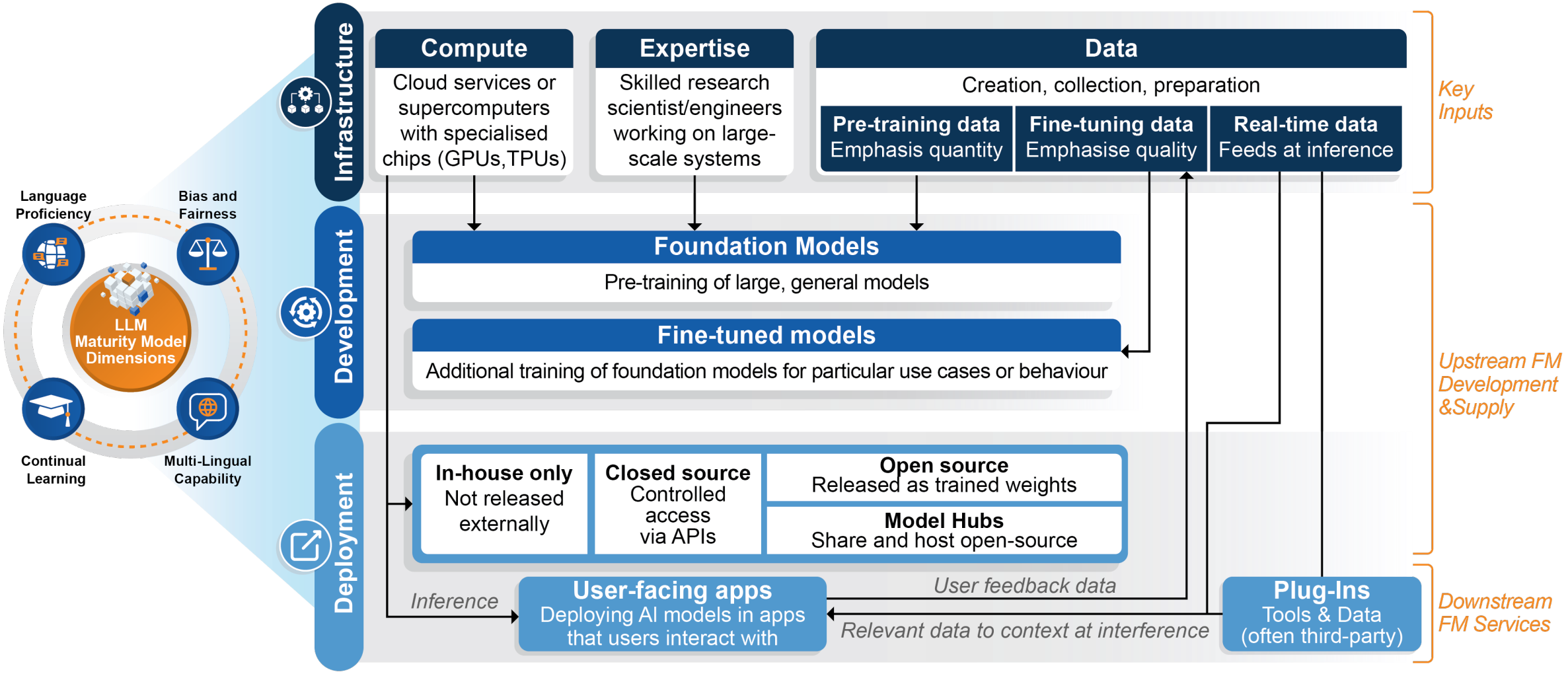
- *User Interaction*
- *Data Incorporation*
- *Fine-Tuning*
- *Version Control*
- *Model Evolution*
- *Ethical Considerations*

Multi-Lingual Capability

- *Linguistic Proficiency*
- *Language Coverage*
- *Cross-Language Transfer*
- *Language Pairs*
- *Cultural Sensitivity*
- *Localization*

Evaluate and measure LLM Maturity across 4 dimensions and 24 representative sub-dimensions

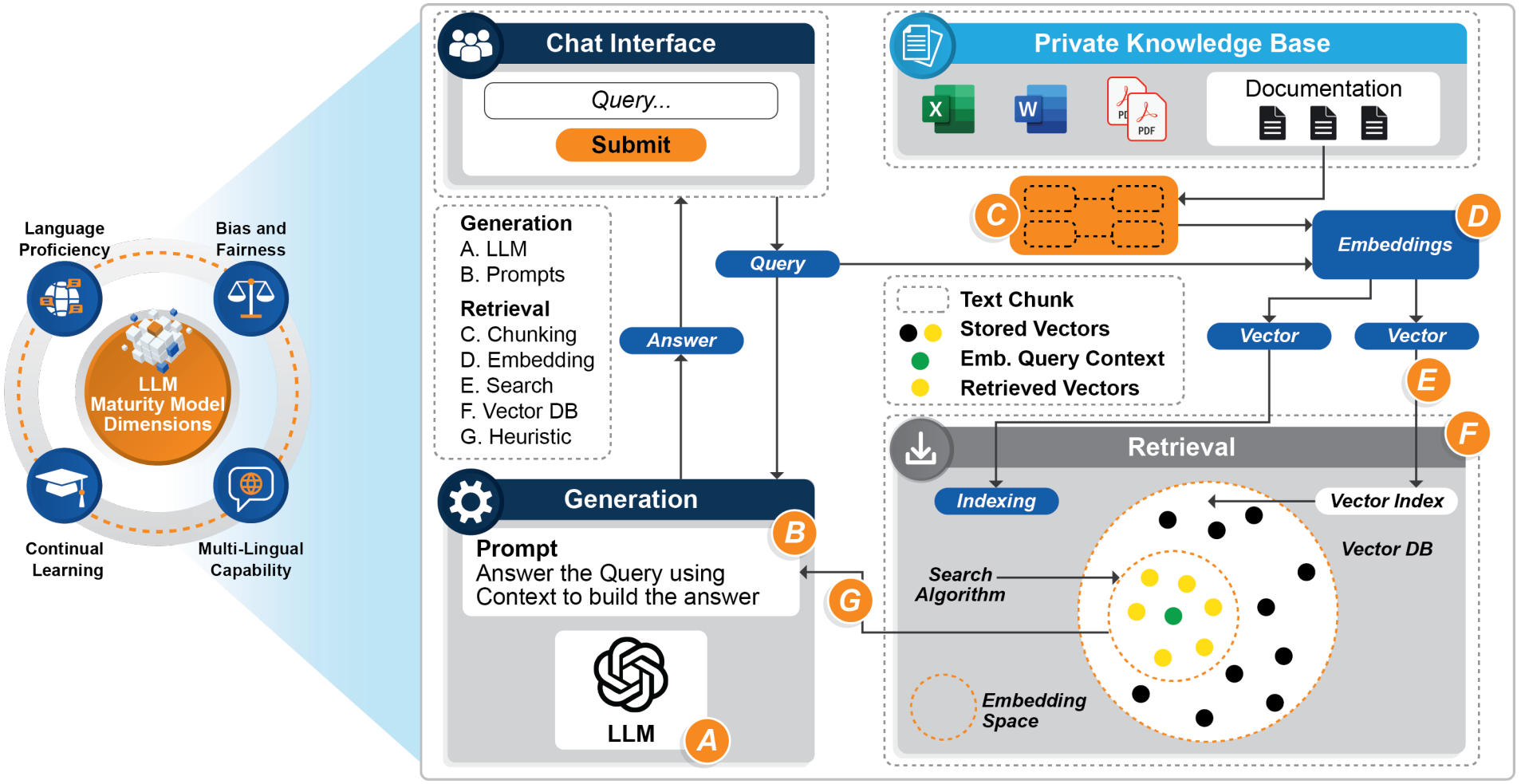
LLM Model Evaluations: Infrastructure, Development & Deployment



Source: Competition and Markets Authority, AI Foundation Models Review (2023): https://assets.publishing.service.gov.uk/media/65045590dec5be000dc35f77/Short_Report_PDFA.pdf

Check and evaluate LLM Architectural components across infrastructure, development and deployment to assess current and sustained abilities

LLM Model Evaluations: Use Case Execution Workflow Example



Also check and evaluate across user or business workflows that need to consistently deliver on key LLM use cases

LLM MM Scoring Rubric

Rubric for LLM MM Assessment / Ranking across the four critical dimensions

Evaluation Metrics	Very Low	Low	Satisfactory	High	Very High
Language Proficiency					
Bias and Fairness					
Multi-Lingual Support					
Continual Learning					

Notional scoring rubric - not necessarily needing to be "Very High" for every chosen use case

LLM Maturity Models – Evaluation by Use Case

Model: ChatGPT

USE CASES:

Conversational Chat

Consistency
- Language Proficiency
- Bias and Fairness
- Multi-Lingual Capability
- Continual Learning

Completeness
- Language Proficiency
- Bias and Fairness
- Multi-Lingual Capability
- Continual Learning

Accuracy
- Language Proficiency
- Bias and Fairness
- Multi-Lingual Capability
- Continual Learning

Equitability
- Language Proficiency
- Bias and Fairness
- Multi-Lingual Capability
- Continual Learning

	Score	Score
• *Accuracy and Fluency*	1	4
• *Semantic Understanding*	5	3
• *Multilingual Competence*	3	3
• *Contextual Appropriateness*	5	5
• *User-Friendly Interfaces*	5	4
• *Fine-Tuning*	2	4
• *Bias Detection and Mitigation*	5	5
• *Ethical Guidelines*	5	5
• *Data Representativeness*	5	5
• *Continuous Monitoring*	5	5
• *User Feedback Mechanisms*	5	5
• *Interpretability and Transparency*	5	5
• *Linguistic Proficiency*	1	3
• *Language Coverage*	1	3
• *Cross-Language Transfer*	1	3
• *Language Pairs*	1	3
• *Cultural Sensitivity*	1	3
• *Localization*	1	3
• *User Interaction*	3	4
• *Data Incorporation*	3	4
• *Finetuning*	3	4
• *Version Control*	3	5
• *Model Evolution*	3	2
• *Ethical Considerations*	3	4

Model: ChatGPT

USE CASES:

Summarization

Accuracy
- Language Proficiency
- Bias and Fairness
- Multi-Lingual Capability
- Continual Learning

Completeness
- Language Proficiency
- Bias and Fairness
- Multi-Lingual Capability
- Continual Learning

Governance
- Language Proficiency
- Bias and Fairness
- Multi-Lingual Capability
- Continual Learning

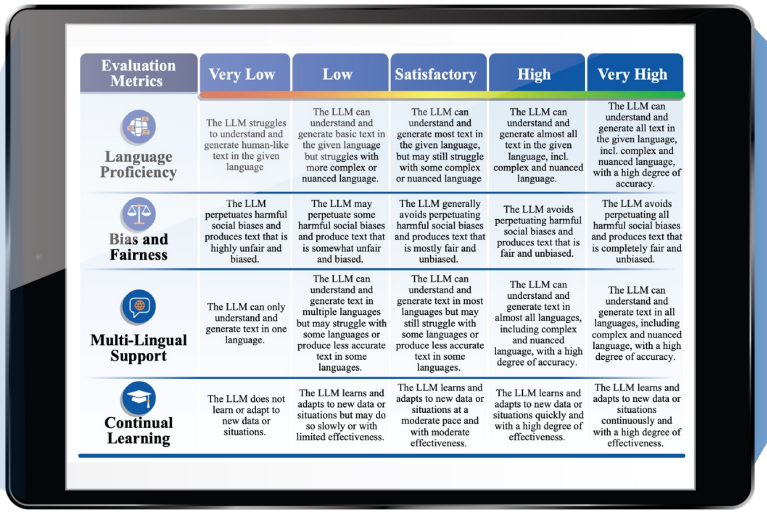
Equitability
- Language Proficiency
- Bias and Fairness
- Multi-Lingual Capability
- Continual Learning

Example use of the rubric across chosen use cases

Heatmap – LLM Characteristics

Use Case	Conversation/ Chatbot
Governability	
Completeness	
Accuracy	
Equitability	
Consistency	
Traceability	
Novelty	
Fluency	
Interactivity	
…	

LLM Characteristics

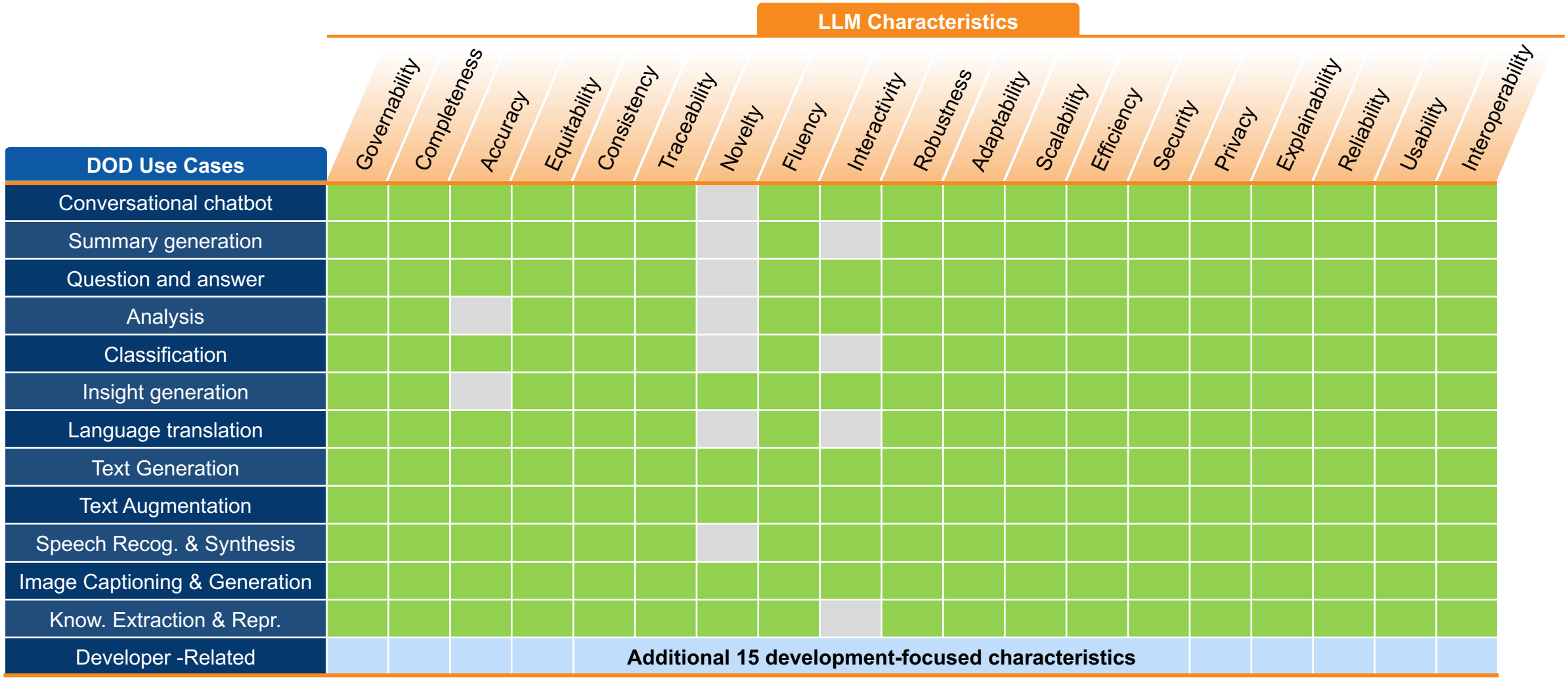


Not of significance to this use case

Use Case	Summarization
Governability	
Completeness	
Accuracy	
Equitability	
Consistency	
Traceability	
Novelty	
Fluency	
Interactivity	
…	

Example evaluation of use cases versus necessary characteristic

Reference Heat Map for Evaluation



Developer–related use cases have an additional set of 15 development-focused characteristics (ex. code generation, summarization, etc.)

Q&A

Contact: srini@iworkscorp.com
srini@computer.org