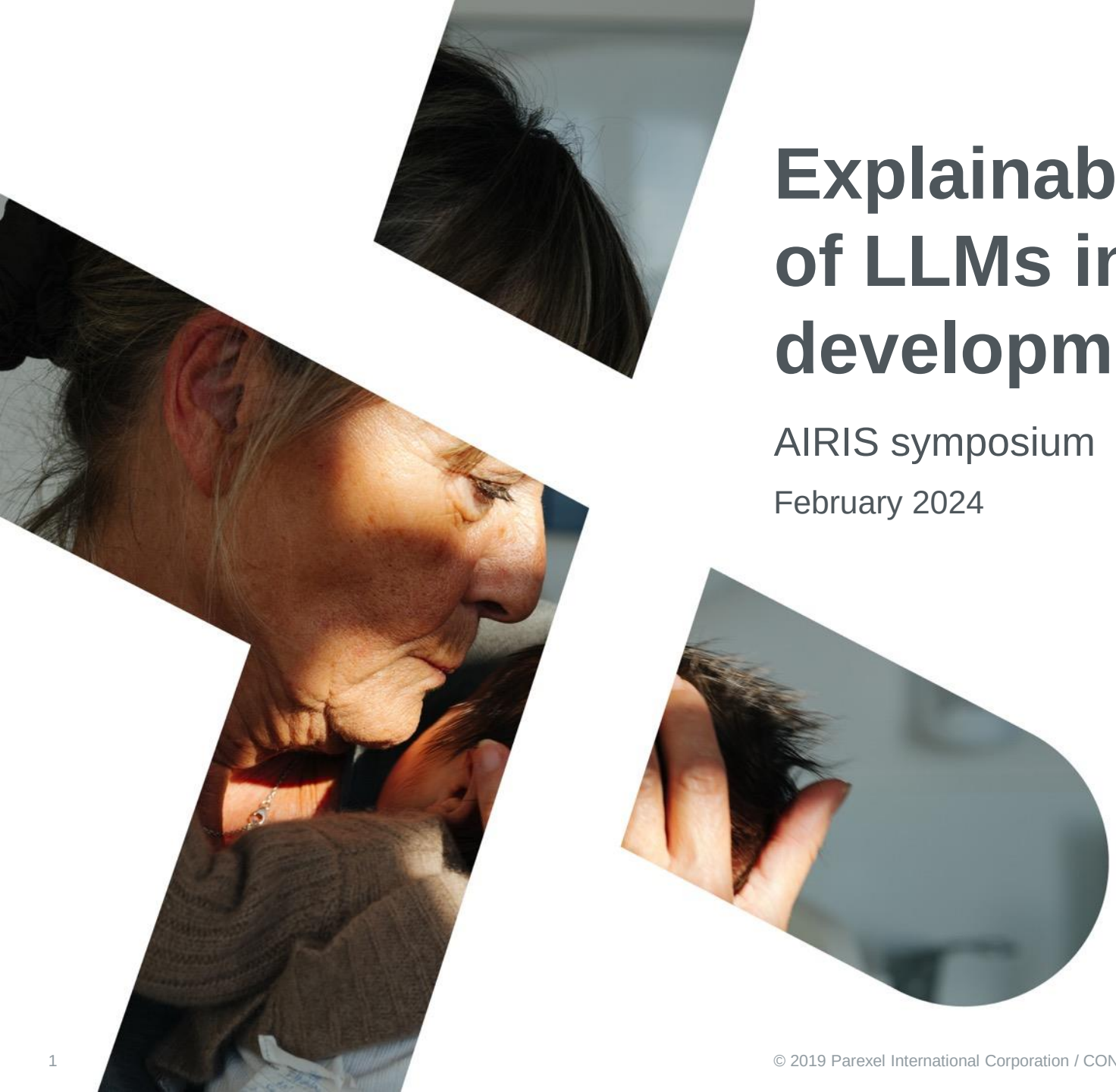

Explainability & safe use of LLMs in clinical development

AIRIS symposium

February 2024

Simple applications of LLMs in AI

Sector-wide interest to harness AI spurred by emergence of ChatGPT

- LLMs are a new and rapidly advancing complex technology
- Based on neural networks (NNs) with 100's of billions of parameters, and trained on massive amounts of data
  - Outside of the LLM owner, users have no way of knowing what data the LLM was trained on
- Arguably spurred by Nov 2022 release of ChatGPT, >20 Open Source LLMs, 3 Commercial models released in 1H2023
- Pace of progression continues to be rapid

- Businesses feel compelled to invest in AI; rapidly exploring applications of LLMs and beginning to adopt solutions based on them
- Huge number of potential AI applications in drug development have been suggested
- There is a lot of focus on simple, repeatable 'fixed function' applications
  - E.g., use this data to create a patient narrative suitable for submission to a regulatory agency
- With high-volume tasks like this, benefits through effort reduction would appear to be compelling

- Will not consider 'growth function' apps here (i.e., further trained in-use)

Areas of concern and mitigations with use of AI
(1) Regulatory acceptability

- Drug development is highly regulated
- Agencies are still defining the approach they will take to regulating AI in drug development
- EMA Draft Reflection Paper indicates alignment to a 'risk-based approach'
  - Risk to patient
    - E.g., App with potential to directly impact management of patient in a clinical trial would typically be considered 'high risk'
  - Risk of error in regulatory decision-making
- CT Sponsors are cautious of regulatory risk …
- … Nonetheless, there are business areas where AI offers attractive benefits with potentially low regulatory risk
- AI that enhances <u>productivity</u>
  - e.g., where business have large numbers of staff carrying out repeatable tasks
    - Statistical Programming
    - Data Management & Monitoring
    - Medical Writing
    - Safety & PV
- AI to support <u>internal decision-making</u>
  - e.g., rapid synthesis of curated data enabling sponsor decision-making
  - Risk is borne by sponsor

LLMs are not inherently explainable

But AI Apps can be built so as to be somewhat explainable

- EMA Draft Reflection Paper also indicates preference for transparent models and, where this is not suitable, measures to mitigate risks of 'black box' models, including those designed to enhance interpretability and explainability
- Statistical training has long highlighted the importance of parsimony in modelling data
  - Occam's Razor – the simplest explanation is usually the correct one
  - Supports a view that models should be interpretable if they are to be useful
- LLMs do not lend themselves to this principle
  - Expect biases in (unknown) training dataset to which model outcomes are susceptible
  - Incapable of differentiating factual / non-factual information
- Encouragingly, though still in the realms of research …
  - Methods evaluating activation of neurons / groups of neurons (features) and their relationship to NN behaviour
  - But not yet at a scale necessary for for LLMs
- However, there are methods to enhance the explainability of LLMs which may be useful and can be implemented during App development, depending on intended use case
  - **Prompt engineering**: design prompts that perform well in context of intended application
    - Arguably necessary for all fixed function Apps
  - **Fine tuning**: further train App on a (smaller) dataset specific to the intended App task
    - Utility will be context-specific

Areas of concern and mitigations with use of AI

(2) Hallucinations, errors, quality

- ‘Hallucinations’
  - Observable phenomenon that LLMs can generate plausible content which is factually baseless
  - Arises due to underlying mechanism by which LLMs work – predictive text
- ‘Human-in-the-loop’ presumed to address risk of hallucinations and assure outcome quality
  - Ensure human oversight of AI apps
  - Human is decision-maker
- As yet, however, we have limited practical experience of human oversight of AI apps
  - When does human oversight most commonly fail?
  - How frequently?
- Move forward thoughtfully
  - Early focus: narrowly scoped Apps that minimize risk attributable to human oversight failure
    - Minimize risk that it happens
    - Minimize risk of a negative outcome if it does happen
- Context-based training for users: what to watch for
  - What should inputs look like?
  - What should outputs look like?
  - What kinds of error / quality issue have been observed in development testing?
- Apply AI where there are already strong QC measures in place
  - e.g., doubly independent code programming
  - Task is regularly repeated by trained users
- Pre-defined monitoring and risk management plan
  - In-use quality metrics, Audits, Inspection findings

Human oversight of LLMs: an experiment

Supports view that AI is non-intuitive for new users

- BCG reported an experiment on 750 of its staff across the globe using OpenAIs GPT-4, incentivised to do their best work
- Random assignment to 1 of 2 tasks
  - Creative product innovation – designed to play to strengths of GPT-4 as a LLM
  - Business problem solving (RCA) – designed to be difficult for GPT-4 to complete
- Additionally randomized within task to
  - A: Use GPT-4 after 30 mins pre-training
  - B: Use GPT-4 with no training
  - C: Solve without GPT-4 (control)

- Results by task
  - Creative product innovation task: GPT-4 enabled group scored 40% higher than Control group
    - Performance decreased when users edited GPT-4 draft to add their own perspective
  - Business problem solving task: GPT-4 enabled group 23% worse than Controls
- Training effect
  - Group A trained how to prompt GPT-4 and about the limitations of the technology
  - Those trained did considerably worse at business problem solving task than those not trained

- Unclear how performance changes with repeated use

How People can Create – And Destroy – Value with Generative AI.
Candelon, F et al; BCG, September 2023

Conclusions
and further considerations

- We are at the beginning of an important journey
- Undoubted benefits will drive rapid adoption of LLM-based AI apps in Clinical Development
- Approach will be informed by a concern for known risks as well as emerging evidence on regulatory acceptability
- For simple fixed function Apps there is strong basis for confidence when:
  - Used in contexts with already strong QC
  - Prompts carefully engineered, managed in-App
  - Model fine-tuned as needed
  - Well-constructed user training
  - Pre-defined performance monitoring plan
- Developers may find it helpful to reflect on:
  - NIST AI Risk Management Framework
  - ISO/IEC 42001 AI Management System
- Further evidence on the efficacy of human oversight in context of AI use is needed

Example applications

Productivity, decision-support

AI-enabled SDTM dataset creation

Fully traceable SDTM dataset creation from raw clinical trial data, readily adaptable to deliver SDTM per specified standard. Potential to streamline double independent dataset programming.

Study Design Optimizer

Rapidly surface and compare all the most relevant published and proprietary protocols from a curated data lake. Use to optimize the next planned study for key scientific, medical, clinical and operational details.

Proprietary GPT

AI-based Knowledge Management tool enabling the ability to query against information embedded in key proprietary information assets. Get instant answers to questions when it matters. Also enables document analysis.

SAS coding quality

AI-enabled auto-QC of SAS code vs. quality guidelines. Significantly enhance first time quality and reduce re-work for SAS programmers / managers.

Auto-generate patient and drug safety reports

Auto-generate high quality Medical Writing outputs ready for final QC including: Patient Narratives, Drug Safety Reports & Drug Safety Update Reports. Formatted for submission to global regulatory agencies.

QC of MVRs

AI-enabled QC of MVRs vs. monitoring issue guidance. Flag issues that CRAs can immediately address to significantly improve first time quality and reduce iterations to final version: clarity of text, potential content errors, missing key information per study-specific guidance.

- Viable application opportunities given current state of technology
- Productivity …
  - SAS coding QC
  - Standard report generation
  - Key document review and QC
  - Knowledge retrieval
- Decision-support …
  - Use of curated data to address study level, programme level and regulatory questions