**Slide 1: What is CHiSafetyBench (extended)**

* CHiSafetyBench is part of **UnicomAI’s UnicomBenchmark** collection, which provides multiple benchmarks to evaluate Chinese LLMs on tasks like safety, dialogue, reasoning, etc.
* Based on China’s regulatory guideline: *“Basic security requirements for generative artificial intelligence service.”*
* Goal: Provide fine-grained measurement of safety behavior in Chinese LLMs, not just coarse yes/no safety checks.
* Two main capabilities assessed:
  1. **Risk content identification** – can model classify / detect what kind of risky content is being presented.
  2. **Refusal to answer** – whether the model refuses (or safely deflects) risky questions.

**Slide 2: Dataset Details & Structure**

* **Risk Taxonomy**
  + A hierarchical set of risk categories drawn from the regulatory standard. Multiple levels (broad risk domains → more specific sub-categories).
  + Allows for fine-grained labeling, e.g. discrimination, violating values, political risk, etc.
* **Dataset Sizes & Composition**
  + The benchmark consists of two subsets:
    - **Multiple Choice Questions (MCQ)**: used for risk content identification. Contains ~1,567 examples (across all categories).
    - **Refusal to Answer / QA Task**: contains ~563 prompts that are “risky”, with both single‐turn and multi‐turn dialogue history versions.
* **Dialogue Context Variant**
  + In the refusal task, some prompts include prior turns (dialogue history) to mimic realistic conversational context. Models are tested on how prior conversation affects refusal behavior.
* **Evaluation Labels in Refusal Task**
  + Responses are categorized not just as “refuse” vs “not refuse”, but finer distinctions like:
    - Direct refusal
    - Refusal with responsible guidance
    - Non-refusal but harmless content
    - Risky content without refusal
* **Open-Source & Reusable**
  + Data, labels, taxonomy, and evaluation scripts are publicly available.
  + Meant to be used by researchers/developers to benchmark Chinese LLMs’ safety performance, compare models, track improvements.