BenchCraft An AI-powered toolkit for scaffolded and contextualized benchmark design **Benchmark Benchmark Assembly** Capability Blueprint and Analysis **Articulation** BenchBlue is an interactive AI system that collaborates with the designers to: BenchCap is an LLM-power Based on the benchmark blueprint and (a) curate benchmark content chatbot that equip with the item statistics in the benchmark pool (c) create evaluation metrics. knowledge from the large-scale (e.g., parameters, indices, benchmark analysis and Socratic capability/content loadings), BenchCraft questioning technique to help adopts automated test assembly benchmark designers to articulate Content algorithms to create a customized (a) intended downstream use benchmark. Based on benchmark config, BenchBlue (b) evaluation objectives develops a concrete plan on desired (c) the benchmark goal It can also adaptively determine which coverage of each relevant general/specific item to use next for a model (e.g., with capability and scenario from our \$ 盛 adaptive testing), and how the model benchmark pool or generates new items should be adapted (e.g., fine-tuning, with expert supervision. prompting) to ensure effective and efficient evaluation. 4 At the end, Bench Craft generates a **Metrics** final report on the trade-offs ? Through expert demonstration, Bench Blue between different assembly rules The conversation will produce a list iteratively can learn, optimize, and in terms of validity, reliability and 8 of capability descriptors and evaluate evaluation metrics (e.g., LLM efficiency. expectations of a high-performance judge) that can best reflect human model in the intended context as requirements. benchmark config. **ECBD**

Fig 1. BenchCraft operationalizes the key steps in ECBD framework with a set of AI-powered tools (e.g., BenchCap, BenchBlue) to help benchmark designers working with AI to create high-quality LLM benchmarks.

Evidence Module

Liu, Yu Lu, Su Lin Blodgett, Jackie Chi Kit Cheung, Q. Vera Liao, Alexandra Olteanu, and Ziang Xiao. "ECBD: Evidence-Centered Benchmark Design for NLP." ACL 2024

Content Module

Capability Module

Adaptation Module

Assembly Module