TransportationBench: One-Page Summary

Proposed R&D Project and Objectives: TransportationBench creates the first standardized test to measure how well generative artificial intelligence (AI) models understand transportation engineering. The project has three main goals: (1) help transportation agencies trust AI for technical work, (2) make it cheaper and easier to evaluate AI models, and (3) create a testing framework that can adapt as AI technology improves.

Technical Landscape and Problem Addressed: Transportation agencies could greatly benefit from generative AI to help plan roads, analyze traffic safety, and manage transportation systems. However, agencies are hesitant to use AI because they cannot determine if AI models truly understand transportation concepts or might provide incorrect information for critical decisions. Currently, each agency tests AI models on their own, leading to inconsistent methods and wasted effort across different organizations. Existing AI tests focus on general topics like math problems or computer programming, which do not reveal whether AI understands transportation-specific knowledge. Other professional fields have created specialized tests—lawyers have LegalBench to test AI legal knowledge—but no equivalent exists for transportation professionals. This gap prevents agencies from confidently using AI for important transportation decisions.

Proposed Technical Approach and Plan: Our 12-18 month development plan uses two testing methods: knowledge tests and practical application tests. The knowledge tests check if AI models understand transportation concepts at the same level as certified professionals like Road Safety Professionals (RSP1). The practical tests evaluate whether AI can apply this knowledge to real-world tasks like conducting safety audits or analyzing traffic data. Stage 2A (4-6 months) focuses on transportation safety. We will work with certified safety professionals to create 75 test questions and scenarios. We will then evaluate three leading AI models (GPT-5, Sonnet 4.1, Gemini 2.5) to see if they can perform safety analysis tasks as well as human professionals. Stage 2B (8-12 months) expands testing to transportation planning and traffic operations. This includes testing whether AI models can work together like a professional team to complete complex transportation studies. We combine automatic scoring for straightforward questions with AI-assisted evaluation for complex scenarios, while using human experts to ensure accuracy and reduce bias in our testing methods.

Commercialization and Transition Approach: We plan to introduce TransportationBench through two pathways. First, we will work with federal agencies like the U.S. Department of Transportation to incorporate the benchmark into government programs and funding requirements. Second, we will partner with individual transportation agencies to demonstrate practical benefits through real projects. We will present successful case studies at major transportation conferences (eg ITE, TRB, AASHTO, ITSA) to encourage broader adoption.

Potential Impact: TransportationBench enables new AI applications including safety audits that find problems traditional methods miss, traffic signal studies that better balance pedestrian safety with traffic flow, and planning studies that consider multiple transportation modes. The benchmark has great potential to help smaller and rural transportation agencies (95% of all U.S. agencies) overcome persistent gaps in resourcing and improve their innovation capacity by up to 10x.

This one-page summary was formed mindfully with AI (FMAI) as an AI-assisted summary of the original Concept Paper.