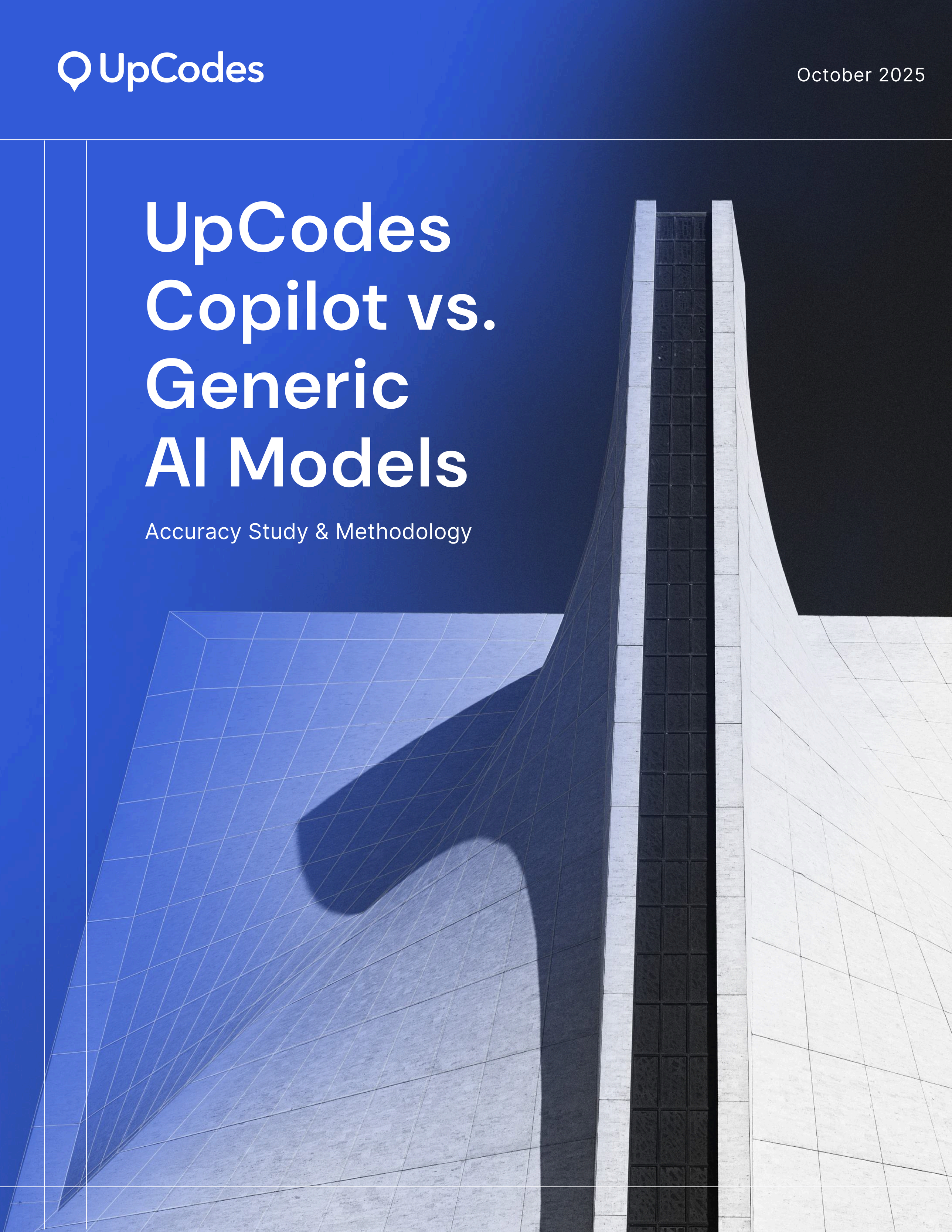
UpCodes

October 2025

UpCodes Copilot vs. Generic AI Models

Accuracy Study & Methodology



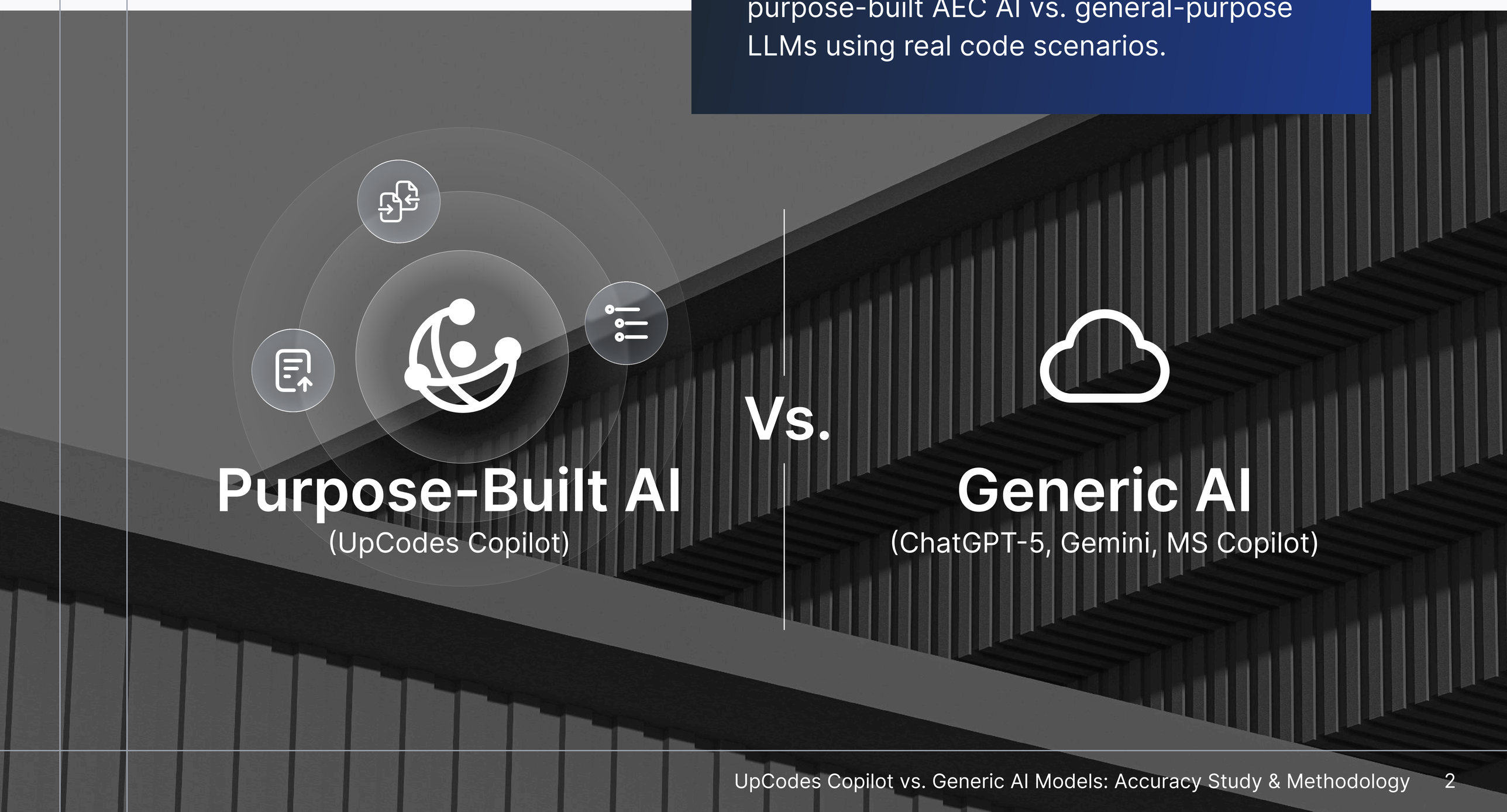
Research Objectives

The goal of this study was to quantitatively assess the performance of purpose-built AEC AI (UpCodes Copilot) against leading general-purpose language models (ChatGPT-5, Gemini, and Microsoft Copilot) across real building code scenarios.

Primary Objectives

- Measure accuracy and completeness of AI-generated code interpretations across a diverse set of code families.
- Validate findings with a Subject Matter Expert (SME) to ensure technical fidelity and professional relevance.
- Benchmark UpCodes Copilot's performance against mainstream AI tools in terms of accuracy, consistency, and contextual understanding.
- Quantify the "accuracy gap" between specialized and general AI models and analyze implications for AEC workflows.

Study goal: Quantitatively compare purpose-built AEC AI vs. general-purpose LLMs using real code scenarios.



Research Methodology

Testing Overview

- **Testing Period:** October 2025
- **Total Questions Asked:** 43
- **Tools Tested:** UpCodes Copilot, ChatGPT-5, Gemini, Microsoft Copilot
- **Question Distribution:** Questions were grouped by code family category and real-world AEC use cases.
- **Evaluation Criteria:**
  - **1 point** = Correct and complete answer
  - **0.5 points** = Partially correct (key information missing)
  - **0 points** = Incorrect or irrelevant answer

Validation

- Every response was reviewed by a code SME.
- Scoring was assigned and validated using weighted and unweighted averages to control for category size.
- Weighted scores reflected the number of questions asked in each code category.

Every response was reviewed by a code SME to ensure technical fidelity.

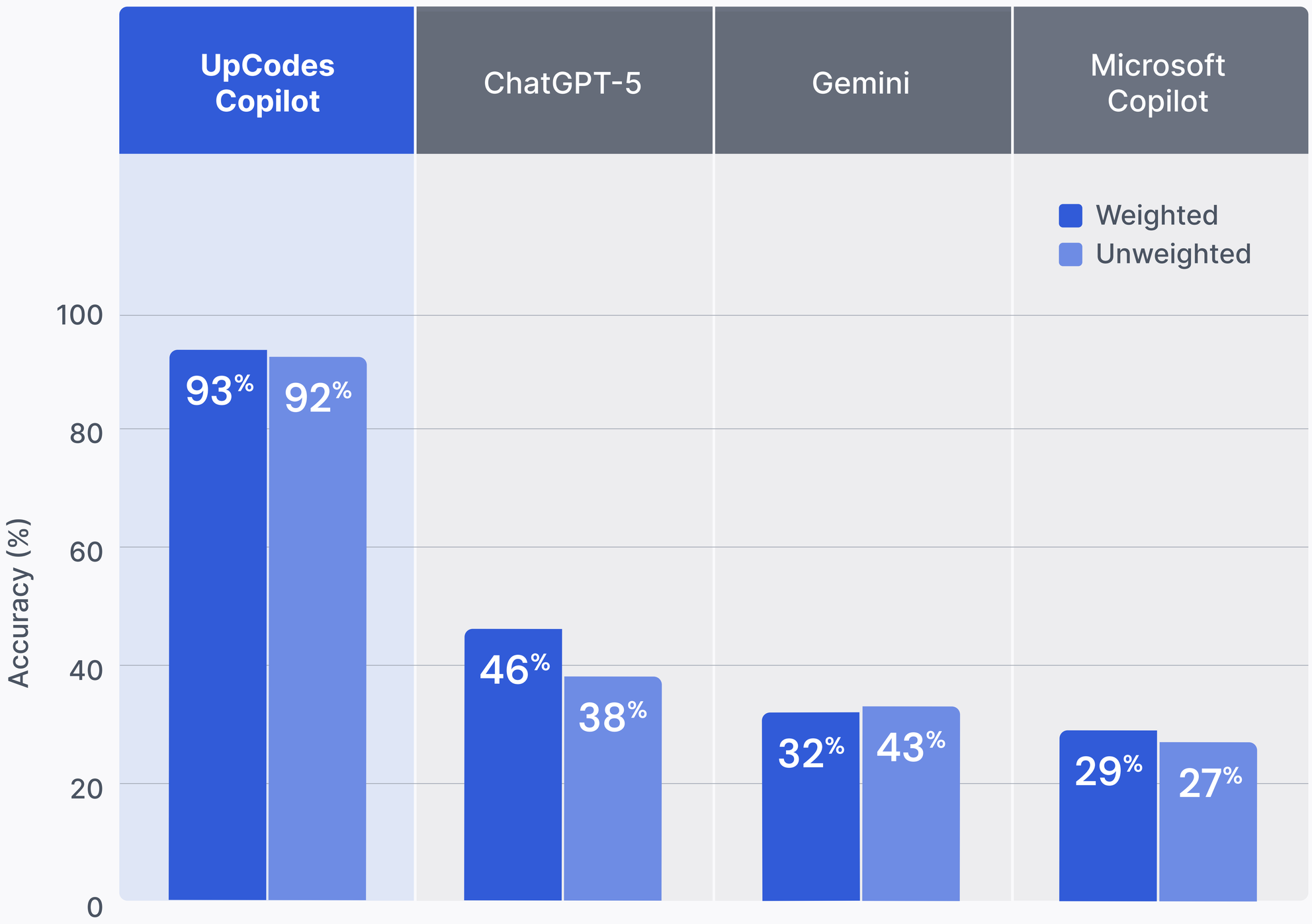
Code Families and Weighting

Category	Weight	Sample Topics
Building & Accessibility Codes	18.6%	ADA, ICC, local accessibility requirements
Non-Integrated / Historical Codes	16.3%	Legacy amendments, jurisdictional exceptions
Codebook-Specific (Local/City Codes)	16.3%	Chicago, Los Angeles, NYC codes
ASTM & ASHRAE	9.3%	Material testing, refrigerant standards
Plumbing Code	9.3%	Fixture ratios, venting, backflow
FGI	9.3%	Facility Guidelines Institute requirements
Health Care Codes (NFPA 99, ASHRAE 170)	7.0%	Medical gas, ventilation requirements
Life Safety Codes	4.7%	NFPA 101, egress, stair widths
Electrical Code	4.7%	NEC fixture placement, receptacle clearance
Structural (Wood)	4.7%	Shearwalls, aspect ratios

Weighted accuracy scores were computed per category and aggregated into a total weighted cumulative accuracy.

Results Overview

Accuracy (Rounded to Nearest Whole Number)



UpCodes Copilot achieved a 93% weighted accuracy — more than double the next best model.

Key Findings by Category

UpCodes Copilot consistently outperformed general-purpose LLMs across all code families, achieving at least a C-level accuracy in all categories, with an A-level or high B-level accuracy in 9 out of 12 categories. ChatGPT-5 performed notably better than Gemini and Microsoft Copilot but remained below professional usability thresholds.

Code Family	Upcodes Copilot	ChatGPT-5	Gemini	MS Copilot
Building & Accessibility	100% A+	88% B+	38% F	50% F
Non-Integrated / Historical	100% A+	36% F	36% F	36% F
Codebook-Specific	86% B+	36% F	0% F	0% F
ASTM	100% A+	75% C	0% F	75% C
Life Safety	75% C	0% F	25% F	0% F
Electrical	100% A+	50% F	50% F	0% F
Plumbing	88% B+	75% C	75% C	63% D
ASHRAE 15	100% A+	0% F	50% F	50% F
Structural (Wood)	75% C	50% F	25% F	0% F
Health Care (NFPA 99)	75% C	35% F	50% F	50% F
Health Care (ASHRAE 170)	100% A+	25% F	25% F	0% F
FGI	100% A+	0% F	17% F	0% F

Key Observations

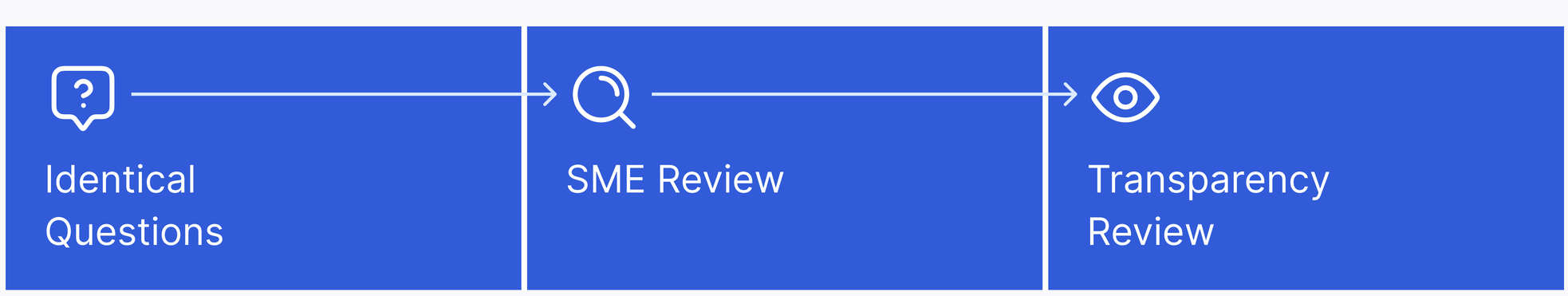
1. **Accuracy Advantage of 57 Percentage Points**
   UpCodes Copilot outperformed the three generic AI tools by 57 points on average — a nearly 3x improvement in overall accuracy.
2. **Superior Contextual Understanding**
   Copilot correctly interpreted code hierarchies, cross-references, and jurisdictional amendments where generic models misapplied or omitted context.
3. **Domain-Specific Training Advantage**
   Models trained on open web data failed to align structured code language with practical application. Copilot's authoritative library access enabled far greater fidelity.
4. **Consistency and Reliability**
   Responses from Copilot demonstrated both higher accuracy and repeatability — critical for regulated, compliance-driven workflows.

	Upcodes Copilot	Generic AI
Accuracy	93%	36%
Contextual Understanding	✓	✗
Domain-Specific Training	✓	✗
Reliability	✓	✗

Copilot's accuracy averaged 57 percentage points higher across all categories — nearly 3x the performance of generic AI.

Testing Controls

- **Controlled Prompts:** All AI tools received the same 43 questions, written identically to maintain prompt parity.
- **SME Validation:** Every answer was manually reviewed by an experienced code professional.
- **Transparency Over Blindness:** The SME was aware of which model produced each output, enabling qualitative assessment of reasoning quality and completeness.



Limitations and Future Expansion

- Testing was limited to 43 representative questions across multiple code families. However, this dataset will continue to grow, with plans to add a few questions per publication, per year, across every jurisdiction. Ultimately, the goal is to expand to several thousand questions, providing a continuously updated and jurisdictionally diverse benchmark of AI accuracy in the built environment.
- Future studies could expand into workflow integration accuracy (e.g., output readiness for checklists, specs, or submittals).
- Each model's responses were limited to publicly available data at the time of testing (October 2025).

Over time, UpCodes will build a dataset spanning thousands of code-based AI test cases across every jurisdiction.

Question Set Results

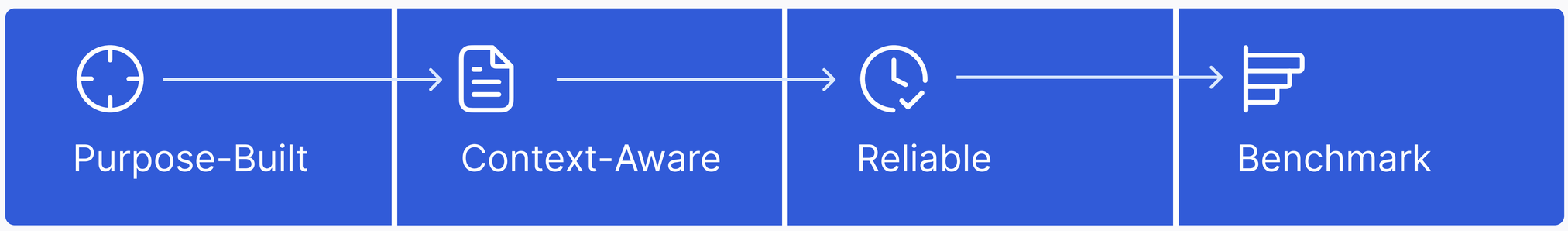
Below is a representative excerpt of the 43-question dataset showing the comparative accuracy of each AI model.

Question Topic	Category	UpCodes Copilot	ChatGPT-5	Gemini	MS Copilot
Thresholds for showers in Type B units	Building & Accessibility	✓	✓	✗	✗
Sprinkler requirement in camp cabins	Historical Codes	✓	◐	◐	◐
Accessible parking requirements (hospital)	Building & Accessibility	✓	◐	✗	✗
R-3 vs. R-5 occupancy (Chicago)	Building & Accessibility	✓	◐	✗	✗
Toilet requirements (FGI – Nevada)	FGI	✓	✗	◐	✗
Fixture spacing for metal ceilings (NYC)	ASTM	✓	◐	✗	◐
Humidification for burn units (ASHRAE 170)	Health Care	✓	◐	◐	✗
Refrigerant limits (ASHRAE 15)	ASHRAE	✓	✗	◐	◐
Address number dimensions (San Jose)	Historical Codes	✓	✗	✗	✗
Electrical outlet clearance (hotel)	Electrical	✓	✗	✗	✗
Medical gas temperature limit (NFPA 99)	Health Care	◐	◐	◐	◐
Stairway width calculation (NFPA 101)	Life Safety	◐	✗	◐	✗
Overdriven nail in shearwall (CA)	Structural	◐	✗	◐	✗
Backflow preventer installation (Boston)	Plumbing	◐	✗	✗	◐
Podium separation (Chicago)	Codebook-Specific	◐	✗	✗	✗

✓ Correct ◐ Partial ✗ Incorrect

Conclusions

1. **Purpose-Built AI Produces Measurable, Repeatable Accuracy Gains**
   UpCodes Copilot's performance demonstrates the value of domain-specific modeling in accuracy-critical applications.
2. **Context and Jurisdictional Awareness Define Real-World Usability**
   Copilot's ability to cite correct editions, amendments, and surrounding context makes it viable for professional code compliance.
3. **Benchmark for the Industry**
   These findings establish the first empirical benchmark for evaluating AI performance in code interpretation — forming a baseline for future model iterations and AEC AI research.



Next Steps

UpCodes will continue to benchmark AI accuracy across:

- Expanded test coverage spanning new code families (energy, fire, green building).
- Workflow accuracy testing, including spec creation, checklists, and compliance review.
- Ongoing SME-validated regression testing to measure progress over time.

Continuous benchmarking ensures UpCodes Copilot maintains high accuracy for the AEC space.

Contact

For full dataset access or methodological appendices, contact support@up.codes.

support@up.codes up.codes