From an information need to an assessed answer

01

48 information needs

24 public-origin adaptations + 24 new scenarios
Written before exposure to Carrier targets

02

666 pooled candidate–question pairs

57 fixed method settings
Two masked relevance assessors: 1,332 judgments

03

12 development → 36 final

Choose baseline and path parameters on development
Freeze selection before final metrics

04

72 evidence-bound answers

Two readers; complementary 3/3 allocation per domain
Each question appears once to each reader

05

144 answer assessments

Two separate graders; roles, conditions and support
Two co-primary endpoints with Holm adjustment

Fresh agent contexts and frozen files reduce leakage; they do not create independent human validation.