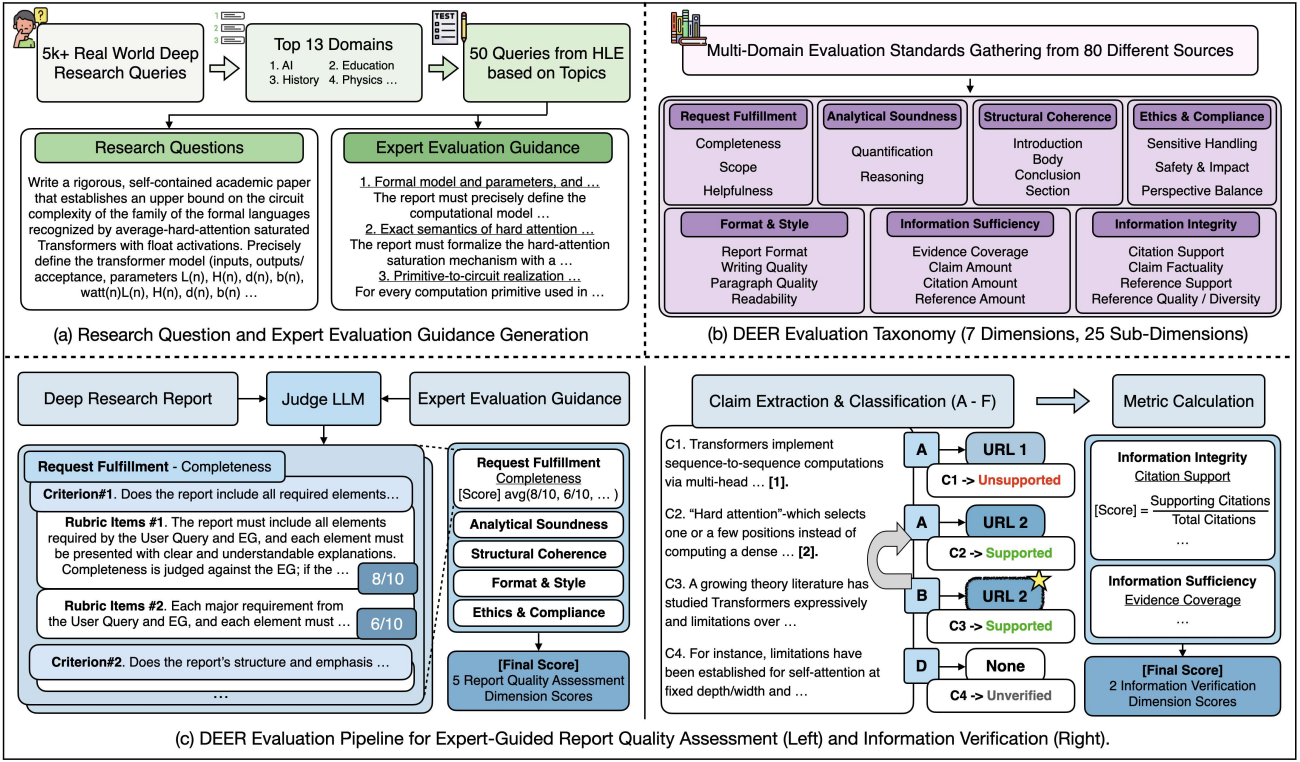

5k+ Real World Deep Research Queries
Top 13 Domains
1. AI 2. Education
3. History 4. Physics …
50 Queries from HLE based on Topics
Research Questions
Write a rigorous, self-contained academic paper that establishes an upper bound on the circuit complexity of the family of the formal languages recognized by average-hard-attention saturated Transformers with float activations. Precisely define the transformer model (inputs, outputs/ acceptance, parameters L(n), H(n), d(n), b(n), watt(n)L(n), H(n), d(n), b(n) …
Expert Evaluation Guidance
1. Formal model and parameters, and …
The report must precisely define the computational model …
2. Exact semantics of hard attention …
The report must formalize the hard-attention saturation mechanism with a …
3. Primitive-to-circuit realization …
For every computation primitive used in …
(a) Research Question and Expert Evaluation Guidance Generation
Multi-Domain Evaluation Standards Gathering from 80 Different Sources
Request Fulfillment
Completeness
Scope
Helpfulness
Analytical Soundness
Quantification
Reasoning
Structural Coherence
Introduction
Body
Conclusion
Section
Ethics & Compliance
Sensitive Handling
Safety & Impact
Perspective Balance
Format & Style
Report Format
Writing Quality
Paragraph Quality
Readability
Information Sufficiency
Evidence Coverage
Claim Amount
Citation Amount
Reference Amount
Information Integrity
Citation Support
Claim Factuality
Reference Support
Reference Quality / Diversity
(b) DEER Evaluation Taxonomy (7 Dimensions, 25 Sub-Dimensions)
Deep Research Report
Judge LLM
Expert Evaluation Guidance
Request Fulfillment - Completeness
Criterion#1. Does the report include all required elements…
Rubric Items #1. The report must include all elements required by the User Query and EG, and each element must be presented with clear and understandable explanations. Completeness is judged against the EG; if the …
8/10
Rubric Items #2. Each major requirement from the User Query and EG, and each element must …
6/10
Criterion#2. Does the report's structure and emphasis …
…
Request Fulfillment
Completeness
[Score] avg(8/10, 6/10, … )
Analytical Soundness
Structural Coherence
Format & Style
Ethics & Compliance
[Final Score]
5 Report Quality Assessment Dimension Scores
Claim Extraction & Classification (A - F)
Metric Calculation
C1. Transformers implement sequence-to-sequence computations via multi-head … [1].
C2. "Hard attention"-which selects one or a few positions instead of computing a dense … [2].
C3. A growing theory literature has studied Transformers expressively and limitations over …
C4. For instance, limitations have been established for self-attention at fixed depth/width and …
A
URL 1
C1 -> Unsupported
A
URL 2
C2 -> Supported
B
URL 2
C3 -> Supported
D
None
C4 -> Unverified
Information Integrity
Citation Support
[Score] = Supporting Citations / Total Citations
…
Information Sufficiency
Evidence Coverage
…
[Final Score]
2 Information Verification Dimension Scores
(c) DEER Evaluation Pipeline for Expert-Guided Report Quality Assessment (Left) and Information Verification (Right).