



## Deep Research Agent



Generated Report



## PDR Dataset & User Settings



Authentic  
Reference  
Materials



User Profile  
& Style Guide



Topics &  
Constraints

### 1. Lexical Overlap Evaluation

Lexical Metrics  
Calculation

Rouge-1

Rouge-L

Meteor

*Content Overlap & Linguistic  
Coherence*

### 2. Quality Evaluation



Generated  
Report



Authentic  
Reference



LLM-as-Judge  
(Pair-wise Score)



1-10

Comprehensiveness (Comp.)

Verifiable, Complete, No Omissions

Readability (Read.)

Language, Structure, Logical Flow

### 3. Personalization Evaluation



Generated  
Report



Authentic  
Reference



LLM-as-Judge  
(Pair-wise Score)



1-10

Content Personalization (C.P.)

Aligns with User Interests, Goals

Presentation Personalization  
(P.P.)

Matches Tone, Style, Formatting

Comprehensive PDR-Eval Score & Analysis  
(7 Metrics)