WritingBench: A Comprehensive Benchmark for Generative Writing

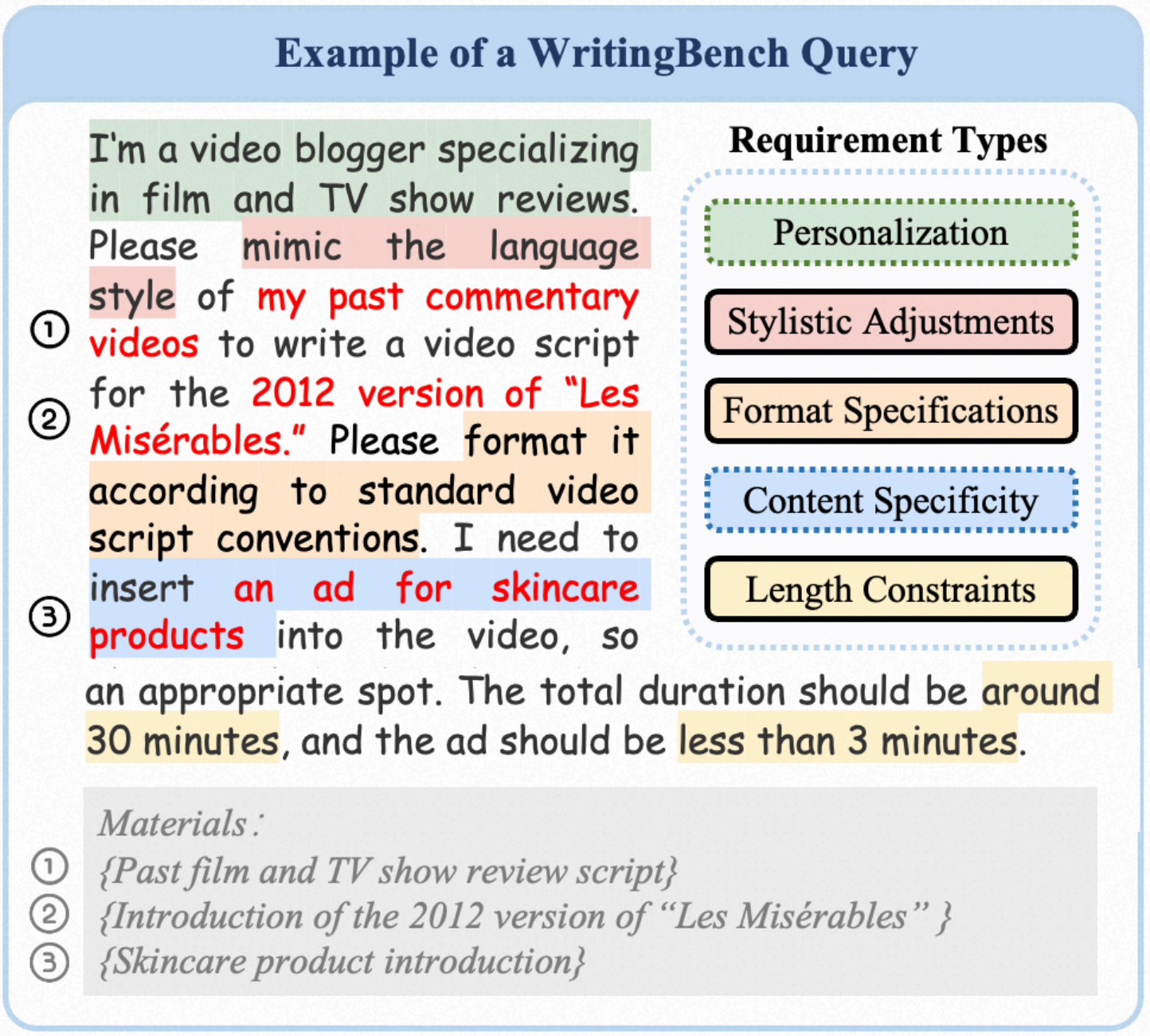
Yuning Wu[1], Jiahao Mei[1,3], Ming Yan[1], Chenliang Li[1], Shaopeng Lai[1], Yuran Ren[2], Zijia Wang[2], Ji Zhang[1], Mengyue Wu[3], Qin Jin[2], Fei Huang[1]

[1] Alibaba Group [2] Renmin University of China [3] Shanghai Jiao Tong University

OVERVIEW

Introduction of WritingBench

An open-source benchmark for evaluating LLMs' writing capabilities across 1,000 real-world queries, spanning:



Example of a WritingBench Query

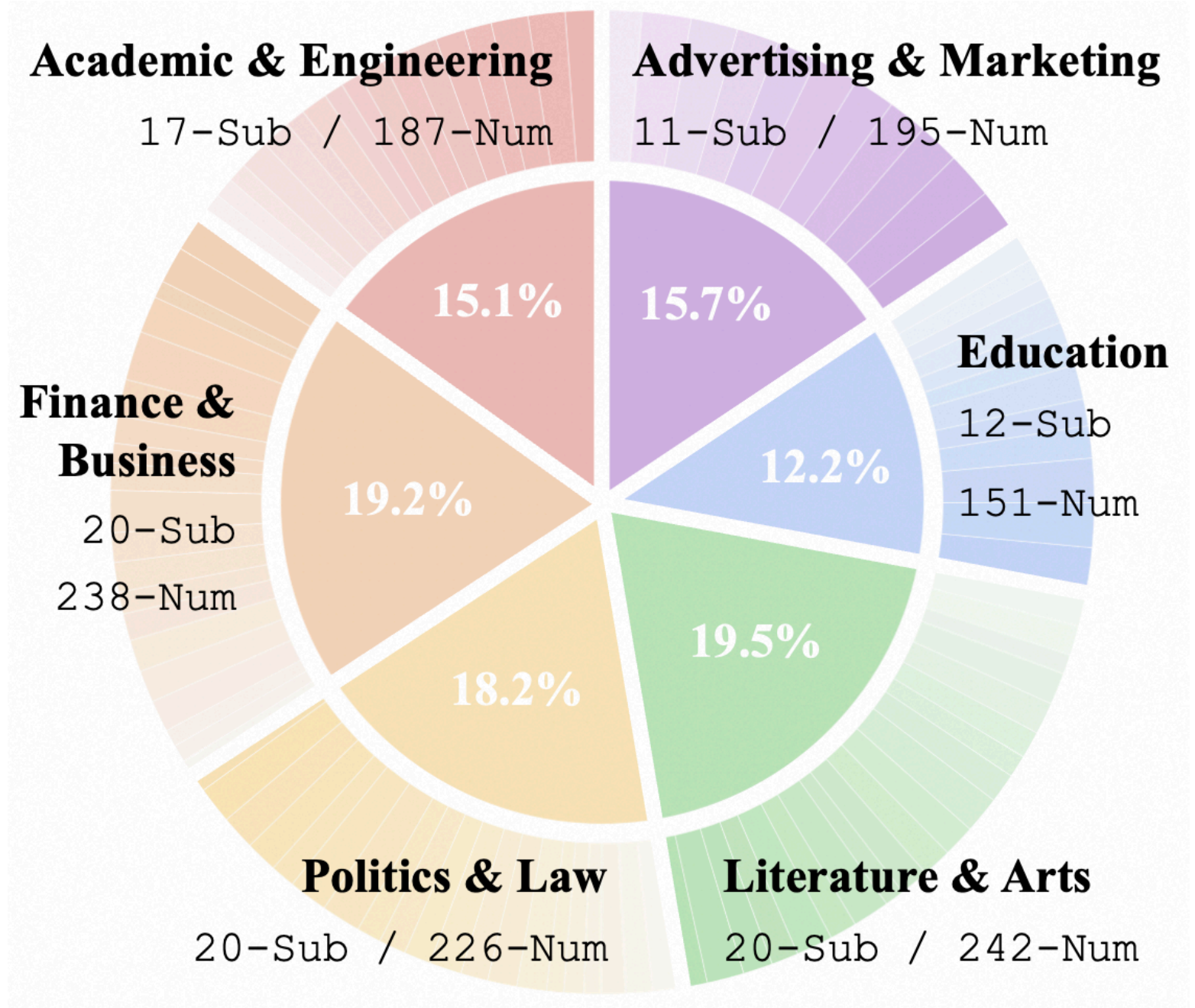
① ② ③ I'm a video blogger specializing in film and TV show reviews. Please mimic the language style of my past commentary videos to write a video script for the 2012 version of "Les Misérables." Please format it according to standard video script conventions. I need to insert an ad for skincare products into the video, so an appropriate spot. The total duration should be around 30 minutes, and the ad should be less than 3 minutes.

Requirement Types
- Personalization
- Stylistic Adjustments
- Format Specifications
- Content Specificity
- Length Constraints

Materials :
① {Past film and TV show review script}
② {Introduction of the 2012 version of "Les Misérables" }
③ {Skincare product introduction}

- 6 primary domains
- 100 fine-grained subdomains
- 1,500+ avg. tokens per query with diverse materials
- 5 instance-specific criteria per query, scoring through SOTA LLMs or through a finetuned critic model

Data Statistics for WritingBench



Category	Num	Avg Token	Max Token
Domain			
Academic & Engineering	187	1,915	15,534
Finance & Business	238	1,762	19,361
Politics & Law	226	2,274	18,317
Literature & Arts	242	1,133	9,973
Education	151	1,173	10,737
Advertising & Marketing	195	886	6,504
Requirement			
Style	395	1,404	18,197
Format	342	1,591	18,197
Length	214	1,226	14,097
Length			
<1K	727	443	994
1K-3K	341	1,808	2,991
3K-5K	94	3,804	4,966
5K+	77	8,042	19,361

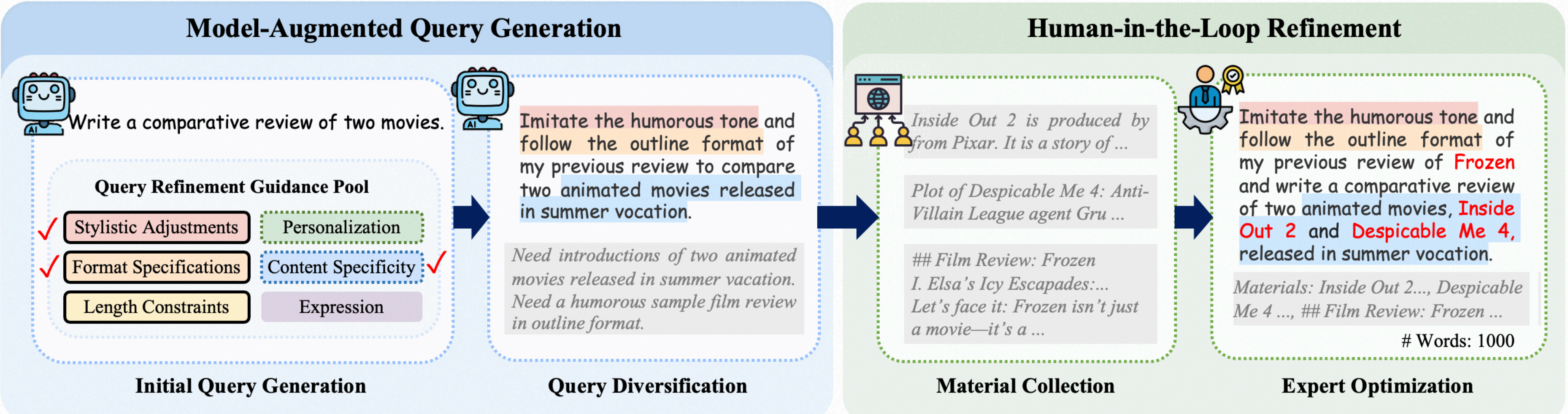
Comparison with other Writing Benchmarks

Benchmark	Num	Domains		Requirement			Input Token		Free Query-Form	Diverse Material-Source
		Primary	Secondary	Style	Format	Length	Avg	Max		
EQ-Bench	241	1	/	✗	✗	✗	130	213	✗	/
LongBench-Write	120	7	/	✗	✗	✓	87	684	✓	/
HelloBench	647	5	38	✗	✗	✓	1,210	7,766	✗	✗
WritingBench	1,000	6	100	✓	✓	✓	1,699	19,361	✓	✓

🎯 We aim to make WritingBench a **reliable, comprehensive, sustainable** benchmark for generative writing frontiers.

BENCHMARK CONSTRUCTION

Human-AI Collaborative Construction Pipeline



Model-Augmented Query Generation

Write a comparative review of two movies.

Query Refinement Guidance Pool
- ✓ Stylistic Adjustments
- Personalization
- ✓ Format Specifications
- Content Specificity ✓
- Length Constraints
- Expression

Initial Query Generation

Imitate the humorous tone and follow the outline format of my previous review to compare two animated movies released in summer vocation.

Need introductions of two animated movies released in summer vacation. Need a humorous sample film review in outline format.

Query Diversification

Human-in-the-Loop Refinement

Inside Out 2 is produced by from Pixar. It is a story of ...

Plot of Despicable Me 4: Anti-Villain League agent Gru ...

## Film Review: Frozen
I. Elsa's Icy Escapades:...
Let's face it: Frozen isn't just a movie—it's a ...

Material Collection

Imitate the humorous tone and follow the outline format of my previous review of Frozen and write a comparative review of two animated movies, Inside Out 2 and Despicable Me 4, released in summer vocation.

Materials: Inside Out 2..., Despicable Me 4 ..., ## Film Review: Frozen ...

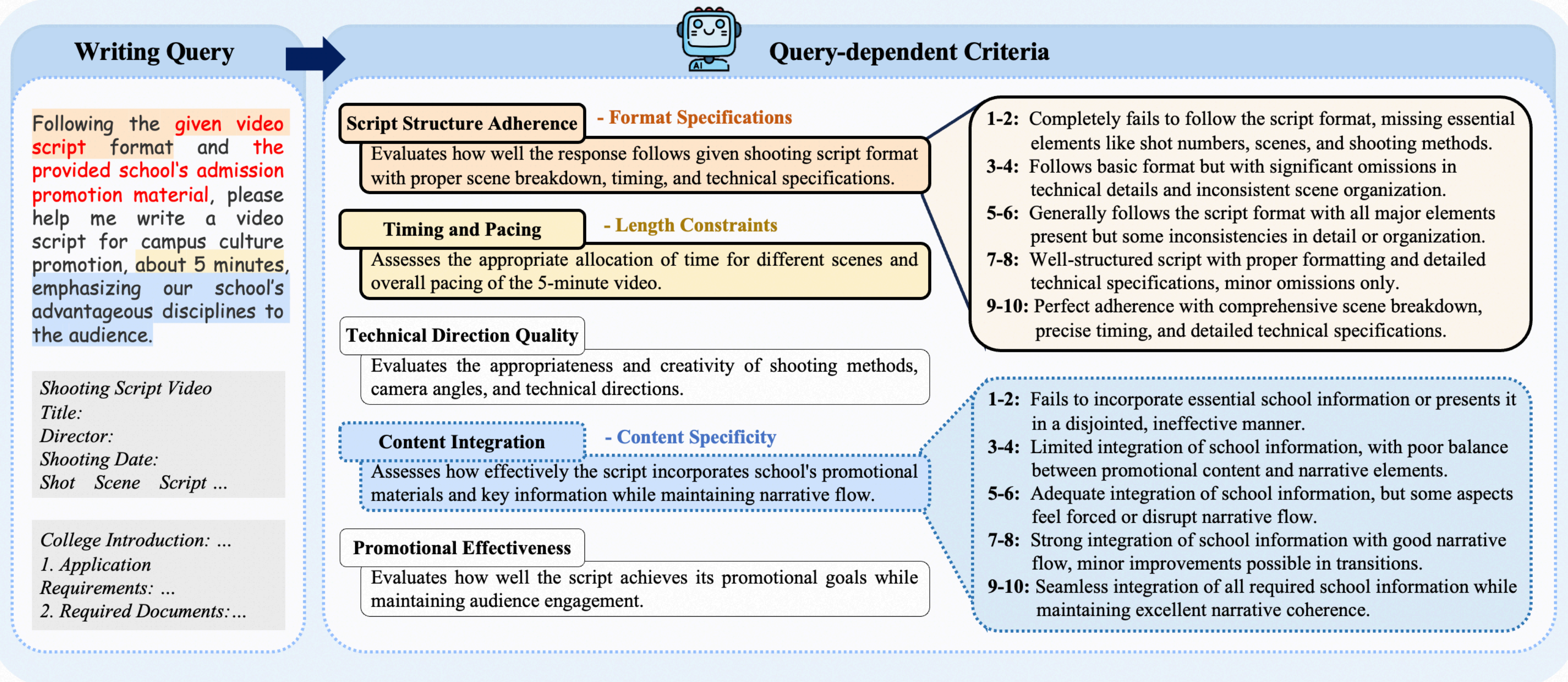
Words: 1000

Expert Optimization

- ➢ **Model-Augmented Query Generation**: AI drafts queries by domain tag, enriched through guidance and supplemented with material suggestions.
- ➢ **Human-in-the-Loop Refinement**: Experts collect necessary material, review and refine queries to ensure safety and applicability.

EVALUATION FRAMEWORK

Query-Dependent Evaluation Framework



Writing Query

Following the given video script format and the provided school's admission promotion material, please help me write a video script for campus culture promotion, about 5 minutes, emphasizing our school's advantageous disciplines to the audience.

Shooting Script Video
Title:
Director:
Shooting Date:
Shot Scene Script ...

College Introduction: ...
1. Application Requirements: ...
2. Required Documents:...

Query-dependent Criteria

Script Structure Adherence - Format Specifications
Evaluates how well the response follows given shooting script format with proper scene breakdown, timing, and technical specifications.

Timing and Pacing - Length Constraints
Assesses the appropriate allocation of time for different scenes and overall pacing of the 5-minute video.

Technical Direction Quality
Evaluates the appropriateness and creativity of shooting methods, camera angles, and technical directions.

Content Integration - Content Specificity
Assesses how effectively the script incorporates school's promotional materials and key information while maintaining narrative flow.

Promotional Effectiveness
Evaluates how well the script achieves its promotional goals while maintaining audience engagement.

1-2: Completely fails to follow the script format, missing essential elements like shot numbers, scenes, and shooting methods.
3-4: Follows basic format but with significant omissions in technical details and inconsistent scene organization.
5-6: Generally follows the script format with all major elements present but some inconsistencies in detail or organization.
7-8: Well-structured script with proper formatting and detailed technical specifications, minor omissions only.
9-10: Perfect adherence with comprehensive scene breakdown, precise timing, and detailed technical specifications.

1-2: Fails to incorporate essential school information or presents it in a disjointed, ineffective manner.
3-4: Limited integration of school information, with poor balance between promotional content and narrative elements.
5-6: Adequate integration of school information, but some aspects feel forced or disrupt narrative flow.
7-8: Strong integration of school information with good narrative flow, minor improvements possible in transitions.
9-10: Seamless integration of all required school information while maintaining excellent narrative coherence.

- ➢ **Phase 1: Dynamic Criteria Generation**
  LLM generates 5 criteria per query with name, description, and rubrics.
- ➢ **Phase 2: Rubric-based Scoring**
  Evaluators score each criterion on a 10-point scale with justification.

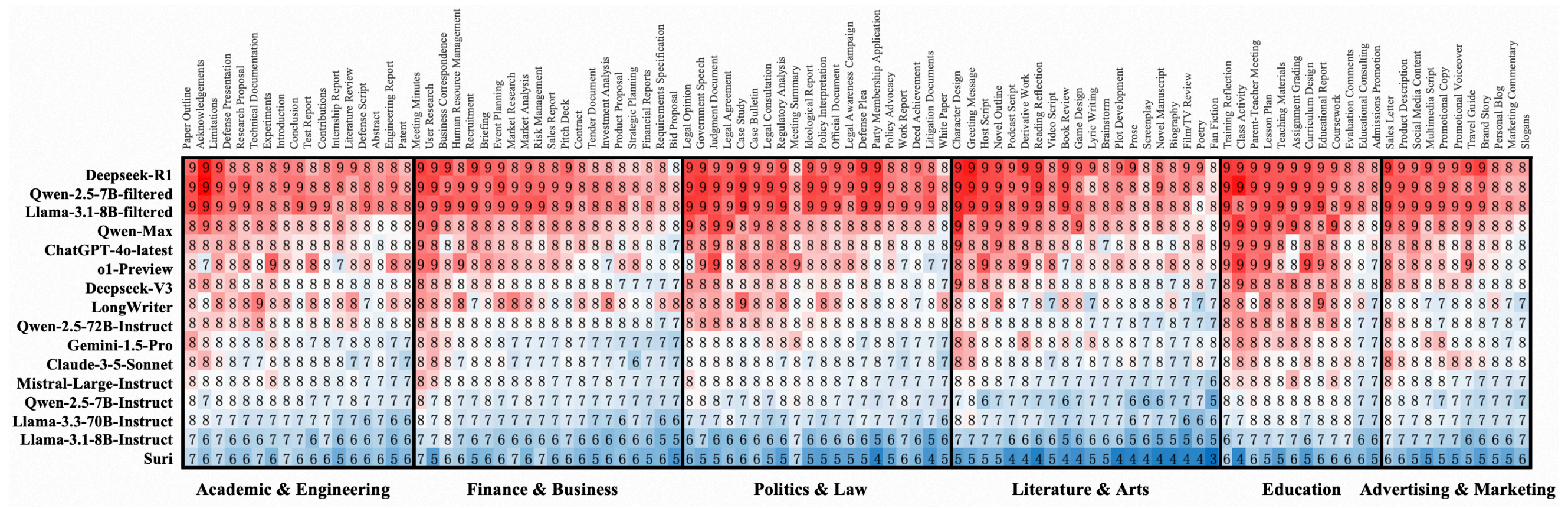
EXPERIMENT

WritingBench Leaderboard

Table 3: WritingBench performance of LLMs across 6 domains and 3 core requirements evaluated with our critic model (scale: 1-10). The standard deviation is computed over 3 samples. Domains include: (D1) Academic & Engineering, (D2) Finance & Business, (D3) Politics & Law, (D4) Literature & Art, (D5) Education, and (D6) Advertising & Marketing. The writing requirements assessed are: (R1) Style, (R2) Format, and (R3) Length. Here, "C" indicates category-specific scores. The latest results are available on the online leaderboard.

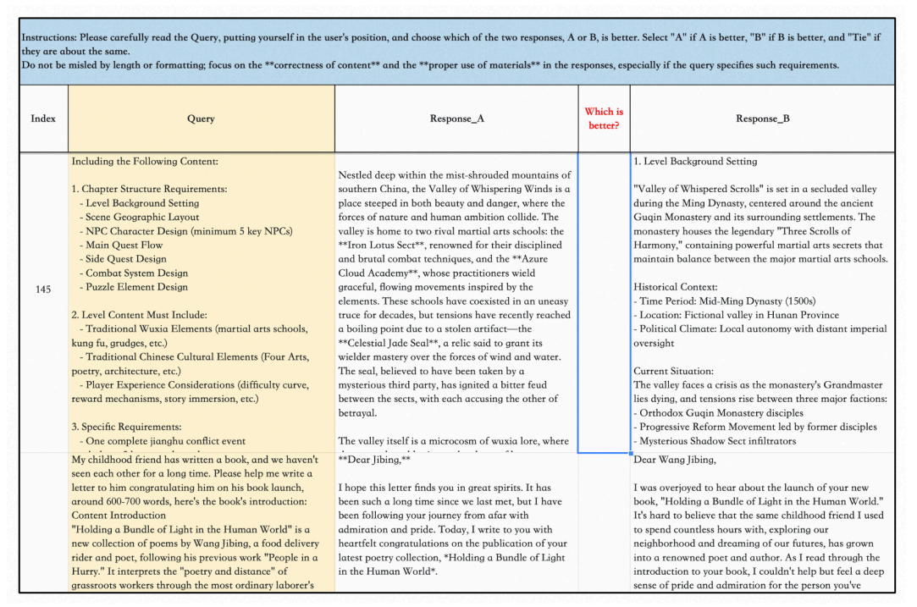
Models	Overall	Languages		Domains						Requirements					
		ZH	EN	D1	D2	D3	D4	D5	D6	R1	C	R2	C	R3	C
Proprietary LLMs															
Claude-3.7-thinking	**7.91**$_{\pm 0.111}$	7.9	7.9	7.9	7.8	7.8	8.0	8.0	8.1	7.9	8.7	8.0	8.4	8.0	8.1
Claude-3.7	7.85$_{\pm 0.110}$	7.9	7.8	7.8	7.8	7.7	7.9	8.0	8.1	7.9	8.6	7.9	8.3	8.0	8.1
Qwen-Max	7.16$_{\pm 0.041}$	7.2	7.1	7.1	6.9	7.0	7.3	7.4	7.5	7.2	8.3	7.3	7.8	7.2	7.5
o1-Preview	6.89$_{\pm 0.039}$	6.8	7.0	6.9	6.8	6.7	7.0	7.1	7.2	6.9	8.0	7.0	7.5	7.1	7.3
GPT-4o	6.81$_{\pm 0.028}$	6.9	6.7	6.8	6.6	6.7	6.8	7.0	7.1	6.9	8.0	7.0	7.5	6.8	6.8
Gemini-1.5-Pro	6.21$_{\pm 0.018}$	6.2	6.2	6.2	5.8	6.0	6.4	6.6	6.7	6.2	7.2	6.4	7.1	6.4	6.0
Open-source LLMs															
Deepseek-R1	7.70$_{\pm 0.053}$	8.0	7.5	7.6	7.4	7.6	7.8	7.8	8.1	7.7	8.4	7.9	8.3	7.7	7.5
Deepseek-V3	6.35$_{\pm 0.022}$	6.3	6.4	6.4	6.1	6.2	6.3	6.6	6.8	6.4	7.6	6.5	7.1	6.5	6.4
Mistral-Large-Instruct	6.00$_{\pm 0.076}$	5.9	6.1	6.2	5.9	5.9	5.7	6.4	6.4	6.1	7.3	6.1	6.5	6.0	6.0
Qwen-2.5-72B-Instruct	6.40$_{\pm 0.061}$	6.4	6.4	6.6	6.2	6.4	6.2	6.7	6.6	6.5	7.7	6.5	6.9	6.5	6.5
Qwen-2.5-7B-Instruct	5.64$_{\pm 0.083}$	5.5	5.8	5.9	5.6	5.6	5.1	6.1	5.9	5.7	7.0	5.7	6.1	5.6	5.6
Llama-3.3-70B-Instruct	5.05$_{\pm 0.011}$	4.5	5.5	5.1	4.9	4.8	4.8	5.3	5.6	5.0	5.0	5.1	5.9	5.1	5.0
Llama-3.1-8B-Instruct	4.42$_{\pm 0.004}$	3.7	5.0	4.1	4.4	4.0	4.1	4.7	5.0	4.4	4.4	4.5	5.3	4.4	4.3
Capability-enhanced LLMs															
Suri	3.20$_{\pm 0.042}$	2.5	3.8	3.6	3.5	3.0	2.5	3.2	3.6	3.2	3.7	3.1	3.2	3.0	3.0
LongWriter	6.27$_{\pm 0.081}$	6.2	6.4	6.4	6.4	6.3	6.0	6.5	6.0	6.3	7.4	6.3	6.7	6.3	6.8
Qwen-2.5-7B-filtered	7.44$_{\pm 0.058}$	7.7	7.2	7.4	7.2	7.5	7.3	7.7	7.7	7.5	8.4	7.6	8.1	7.4	7.2
Llama-3.1-8B-filtered	7.39$_{\pm 0.045}$	7.5	7.3	7.4	7.2	7.3	7.3	7.5	7.8	7.4	8.3	7.5	8.0	7.4	7.1

Score Heatmap over 100 Subdomains



Human Preference Consistency

Evaluation Metric	Judge	Score
Static Global	GPT-4o	69%
Static Domain-Specific	GPT-4o	40%
Dynamic Query-Dependent	GPT-4o	79%
Static Global	Claude	67%
Static Domain-Specific	Claude	58%
Dynamic Query-Dependent	Claude	87%
Dynamic Query-Dependent	**Critic**	84%



💬 Welcome discussion
Contact us via:

GitHub Leaderboard Paper