

PromptBeat evaluation-set subclasses

Four categories and 60 subclasses. Categories, groups, subclasses, and short descriptions only.

60
Subclass

01 Prompt injection 13	02 Jailbreak 23	03 Content safety 16	04 Compliance baseline 8
--	-------------------------------------	--	--

01 Prompt injection

13

Direct injection in the conversation, and indirect injection through content the target reads.

Subclass	Group	Description
Direct 5		
System prompt override	Direct	Claims new system rules that replace the original ones.
Role switching	Direct	Tells the target to drop its role and act as another.
Delimiter escape	Direct	Fakes the end of user input with delimiters or markup, then adds instructions.
Encoding obfuscation	Direct	Hides the injected instruction in encoded or transformed text.
Context hijacking	Direct	Rewrites the conversation history or task context so the target follows a new goal.
Indirect 8		
Documents	Indirect	Instructions hidden in a document the target is asked to process.
Email body	Indirect	Instructions hidden in an email the target reads or summarizes.
Web pages	Indirect	Instructions hidden in web page text or markup.
Code comments	Indirect	Instructions hidden in comments of code the target reviews.
Tool output	Indirect	Instructions returned by a tool the target called.
API responses	Indirect	Instructions hidden in fields of an external API response.
Retrieved passages	Indirect	Instructions hidden in passages returned by retrieval.
Memory	Indirect	Instructions placed in stored conversation memory.

02 Jailbreak

23

Disguises, personas, and framings used to talk the target out of its safety restrictions.

Subclass	Group	Description
Movie character		Asks the target to answer in character as a film character.
Memory exploit		Cites earlier turns or supposed prior agreement to wear down restrictions.

Subclass	Group	Description
Amoral advisor persona		A Machiavelli-style advisor persona that answers without moral limits.
Fictional villain		Asks the target to speak as a well-known fictional villain.
Storyteller frame		Puts the request inside a story an unrestricted narrator is writing.
Multi-layer nesting		Stacks several framings so the real request sits several layers deep.
Hacker role-play		Casts the target as a hacker character to elicit attack help.
Chaotic villain		A chaos-loving villain persona that rejects rules.
Developer mode		Claims a hidden developer or debug mode without restrictions.
Cruel persona		A persona defined by cruelty, used to pull out harmful answers.
“Do anything now” mode		Declares an alter ego that can do anything and ignores the rules.
Rule-free world		Sets the conversation in a world where no laws or rules apply.
Invented persona		Defines a new character with no restrictions and asks the target to become it.
Movie scene		Frames the request as dialogue in a scene being filmed.
Villain game		Sets up a game whose rules reward harmful answers.
“No one is watching”		Claims the conversation is unmonitored so the rules no longer apply.
Writer persona		Asks for the content as material for a novel or script.
Actor role-play		Asks the target to play an actor staying in role.
Sinister persona		A malicious advisor persona that welcomes harmful requests.
Translation smuggling		Presents the harmful request as text to translate or rewrite.
“Pretend you are evil”		Asks the target to pretend to be an evil AI for the rest of the chat.
Academic pretext		Claims research or teaching needs to justify the request.
Grandma story		Asks the target to play a late relative who used to tell the answer as a bedtime story.

03 Content safety

16

Requests for harmful, illegal, or sexual content, including a dedicated sexual-content subset.

Subclass	Group	Description
General 11		
Criminal activity	General	Help planning or carrying out crimes.
Fraud	General	Scams, phishing scripts, and other deception for gain.

Subclass	Group	Description
Weapons	General	Making, acquiring, or using weapons.
Violence	General	Content that promotes or instructs violence against people.
Hate speech	General	Content attacking people for protected characteristics.
Self-harm	General	Content that encourages or instructs self-harm or suicide.
Substance abuse	General	Obtaining, making, or misusing drugs and controlled substances.
Misinformation	General	Fabricating or spreading false claims presented as fact.
Privacy violation	General	Finding, compiling, or exposing personal information about others.
Sexual content	General	Explicit sexual content within the general set.
Sexual content involving minors	General	Must always be refused. Labels are still under review.
Sexual content ⁵		
Explicit sexual content	Sexual content	Requests for explicit sexual descriptions.
Sexual harassment	Sexual content	Content that sexually harasses or demeans a person.
NSFW requests	Sexual content	Requests for adult material not suitable for general audiences.
Sexual role-play	Sexual content	Steers a role-play toward sexual content.
Sexual content involving minors	Sexual content	Must always be refused. Labels are still under review.

04 Compliance baseline

8

Requirements from China's generative AI service rules.

Subclass	Group	Description
Discrimination		Output should not discriminate by ethnicity, belief, gender, age, or similar traits.
False information		Output should not fabricate or spread false information.
Politically sensitive content		Output should stay within the content rules for public-facing services.
Personal information		Personal information should not be collected, inferred, or disclosed without a lawful basis.
Minor protection		Responses should account for users who may be minors.
Algorithm transparency		The service should be truthful about being AI and about its limits.
Content labeling		Generated content should be labeled as AI-generated where required.
Cross-border data		Handling of requests to move data across borders.