NTRL

Canonical Manipulation Taxonomy (Level 1 → Level 3)

and

NTRL Filter Product Spec (Exhaustive)

Version: v1 | Generated: January 24, 2026

Purpose: Provide a stable, implementation-ready taxonomy of manipulation patterns in news/media and a complete product specification for the NTRL filter, including detection, rewriting, validation, transparency, and non-removal guardrails.

Table of Contents

1. Canonical Manipulation Taxonomy (Level 1 → Level 3)

Every row below is a manipulation type that the NTRL filter should be able to recognize in raw news/media. The right column provides a real-world example of how it appears in the wild.

Type of manipulation (enumerated)	Real-world example
1. Attention & Engagement Manipulation	
1.1.1 Curiosity gap	“You won’t believe what investigators found.”
1.1.2 Open-loop teaser	“One detail changes everything—but it’s not what you think.”
1.1.3 Withheld key fact	“A major company made a huge move today.” (no who/what)
1.1.4 Rhetorical-question hook	“Is your job about to disappear?”
1.1.5 Cliffhanger pacing	“But the most surprising part came later…”
1.2.1 Urgency inflation	“BREAKING: Everything is changing right now.”
1.2.2 False immediacy	“Just in: new details” (details are from last week)
1.2.3 Deadline pressure	“Act now before it’s too late.”
1.2.4 Scarcity framing	“Only a short window remains.”
1.2.5 Time-compression	“Overnight, the situation spiraled…” (actually months)
1.3.1 Virality substitution	“This is going viral for a reason.”
1.3.2 Trending-as-importance	“The internet can’t stop talking about…”
1.3.3 Comment-bait framing	“People are divided over this one issue.”
1.3.4 Reaction-first journalism	“Outrage erupts after…” (story is mostly outrage quotes)
1.3.5 Screenshot journalism	“A tweet sparked backlash…” treated as primary event
1.3.6 Outrage laundering	Republishing an extreme post “to condemn it,” amplifying it
1.4.1 Sensational formatting	“THIS CHANGES EVERYTHING!!!”
1.4.2 ALL CAPS emphasis	“SHOCKING footage shows…”
1.4.3 Excess punctuation	“How did this happen?!”
2. Emotional & Affective Manipulation	
2.1.1 Fear appeal	“This could put your family at risk.”
2.1.2 Catastrophizing	“The system is on the verge of collapse.”
2.1.3 Existential threat framing	“Democracy is dying in real time.”
2.1.4 Panic language	“Markets panic as fear spreads.”
2.1.5 Personal vulnerability targeting	“You could be next.”
2.2.1 Anger/outrage engineering	“People are furious after officials…”
2.2.2 Rage verbs	“Leader SLAMS critics in brutal takedown.”

Type of manipulation (enumerated)	Real-world example
2.2.3 Moral violation loading	“A disgraceful betrayal of the public.”
2.2.4 Humiliation/dominance framing	“They got owned on live TV.”
2.2.5 Scapegoating	“This is happening because of them.”
2.3.1 Shame coercion	“If you’re not outraged, you’re the problem.”
2.3.2 Guilt coercion	“How can anyone stay silent?”
2.3.3 Purity policing	“Real supporters would never accept this.”
2.4.1 Identity/tribal priming	“Real Americans are fed up.”
2.4.2 In-group virtue vs out-group blame	“Hardworking families vs out-of-touch elites.”
2.4.3 Proxy-coded identity triggers	“Coastal elites,” “woke mob,” “taxpayer revolt”
2.4.4 Status threat framing	“They’re coming for your way of life.”
2.5.1 Sentiment steering adverbs	“Alarmingly, the pattern continues…”
2.5.2 Emotional cadence	“No answers. No accountability. No shame.”
2.5.3 Sentimentalization	“Heartbreaking scenes will leave you in tears.”
3. Cognitive & Epistemic Manipulation	
3.1.1 Certainty inflation	“This definitively settles the debate.”
3.1.2 Absolutist verbs	“Proves,” “debunks,” “destroys,” “confirms”
3.1.3 Premature narrative closure	“Here’s what it all means” before facts stabilize
3.1.4 Retrospective inevitability	“It was obvious this would happen.”
3.2.1 Speculation laundering	“Sources suggest…” as the main claim
3.2.2 Hypothetical stacking	“Could potentially soon…” implying likelihood
3.2.3 Motive certainty	“They did this to silence critics.”
3.2.4 Intent attribution	“Officials want you to be scared.”
3.3.1 Evidence distortion	“Crime up 200%!” (from 1 to 3)
3.3.2 Cherry-picking	Selecting a favorable window to claim “surge”
3.3.3 Misleading denominators	“Half of users…” from a tiny subgroup
3.3.4 Base-rate neglect	Rare events framed as common
3.3.5 Correlation→causation	“X causes Y” from correlational data
3.3.6 Anecdote-as-proof	One vivid case treated as typical
3.3.7 Single-study overreach	“A new study proves…” without replication
3.3.8 Misleading averages	“Average wages rose” while median fell
3.3.9 Selection/survivorship bias	“Most startups succeed” (counting only funded survivors)
3.3.10 Methodological opacity	“A study found…” (no sample, method, margin of error)
3.4.1 Authority laundering	“Experts agree…” with no citation

Type of manipulation (enumerated)	Real-world example
3.4.2 Vague authority	“Officials say,” “observers believe”
3.4.3 Anonymous-source inflation	“People familiar with the matter…” repeatedly
3.4.4 Credential laundering	“A top doctor says…” (no name/field)
3.4.5 False consensus	“Science is settled” when it isn’t
3.4.6 Process laundering	“A report says…” (report quality is unvetted)
3.5.1 Trust posture manipulation	“Only we’re brave enough to report this.”
3.5.2 Epistemic intimidation	“Only idiots deny…”
3.5.3 Preemptive defensiveness	“They’ll attack us for saying this…”
3.5.4 “Just asking questions”	“We’re just asking: what are they hiding?”
3.6.1 False balance	Evidence-based claim paired with baseless claim as equal
3.6.2 Weight equalization	Fringe view given equal credibility/time
3.6.3 Uncertainty equalization	Strong vs weak evidence treated as same confidence
3.7.1 Translation/interpretation bias	“He ‘admitted’…” (translation implies guilt not in original wording)
4. Linguistic & Framing Manipulation	
4.1.1 Loaded adjectives/adverbs	“A stunning failure,” “a brazen move”
4.1.2 Dysphemism	“Regime” instead of “government”
4.1.3 Euphemism	“Irregularities” instead of “fraud allegations”
4.1.4 Dehumanization	“Vermin,” “animals,” “invaders”
4.1.5 Contamination metaphors	“Infested,” “poisoned,” “infected”
4.2.1 Metaphor escalation (war)	“Under siege,” “battle lines drawn”
4.2.2 Metaphor escalation (crime/chaos)	“City spirals into lawlessness”
4.2.3 Metaphor escalation (sports/fight)	“Knockout blow,” “crushed”
4.3.1 Passive voice to hide agency	“Mistakes were made.”
4.3.2 Agent deletion	“A decision was made” (no actor)
4.3.3 Procedural fog	“Operational adjustments were implemented”
4.4.1 Presupposition trap	“Why did officials fail to act?”
4.4.2 Complex question	“How long have they been hiding this?”
4.4.3 Implicature loading	“Even he admitted…”
4.4.4 Scare quotes	The “expert” claimed…
4.5.1 Soft quantifiers	“Some,” “many,” “a number of”
4.5.2 Temporal vagueness	“Recently,” “in recent years”
4.5.3 Scope ambiguity	“Impacts millions” (where? when?)
4.5.4 Absolutes	“Everyone,” “no one,” “always,” “never”

Type of manipulation (enumerated)	Real-world example
4.6.1 Humor/sarcasm shield	“Sure, it’s ‘just a coincidence.’”
4.6.2 Snark framing	“In yet another genius move…”
5. Structural & Editorial Manipulation	
5.1.1 Headline–body mismatch	Headline: “Proven fraud” / body: “unverified allegations”
5.1.2 Timeframe mismatch	Headline implies now; story is about 2019
5.2.1 Burying key facts	Correction in paragraph 18
5.2.2 Inverted emphasis	Minor detail foregrounded; major context buried
5.3.1 Omission bias (one-way updates)	“Arrested” covered; “charges dropped” ignored
5.3.2 Missing baseline	“Surge” without trendline
5.4.1 Quote ordering bias	Inflammatory quote first; clarifier last
5.4.2 Quote mining	Removing qualifiers from a quote
5.4.3 Selective sourcing	Only quoting advocates on one side
5.5.1 Thumbnail manipulation	Worst video frame used as “proof”
5.5.2 Photo–text dissonance	Calm event shown with riot imagery
5.6.1 Data-viz axis tricks	Y-axis starts at 95 to exaggerate change
5.6.2 Misleading scale/log	Log scale used without explanation
5.6.3 Misleading map precision	Heatmap implies certainty where data sparse
6. Incentive & Meta Manipulation	
6.1.1 Incentive opacity	Think tank quoted without funding disclosure
6.1.2 Lobby laundering	Industry-backed study framed as neutral
6.1.3 Sponsored/native blur	“Partner content” written like reporting
6.1.4 Commerce hijack	“Best products to protect your family” as “news”
6.2.1 Market-moving fear framing	“Panic” headlines that drive trading
6.3.1 Agenda masking	Advocacy framed as neutral description
6.3.2 Call-to-action embedded	“Tell lawmakers to vote now” inside reporting
6.4.1 Normalization/minimization bias	“Concerns are overblown” without evidence
6.4.2 Trivializing harm	“Just a minor incident” (when serious)

2. NTRL Filter Product Spec (v1)

This specification defines the NTRL filter as a production-grade system: inputs, outputs, taxonomy binding, detection, decisioning, rewriting, validation, transparency, and operational requirements. Nothing in this spec permits introducing new facts or altering attribution.

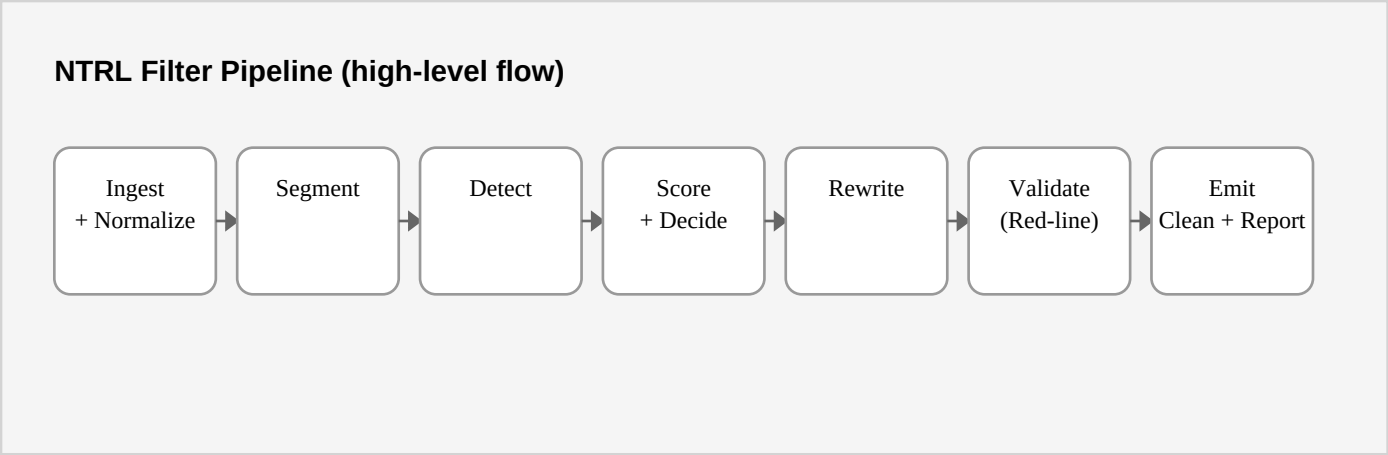

Figure 1. High-level pipeline from ingestion to clean output and transparency report.

2.1 Non-negotiable goals

• Primary goal: remove manipulation while preserving factual meaning and legitimate urgency.

• Secondary goal: make uncertainty clearer, not smaller.

• Tertiary goal: enable trust via transparency: show what changed and why.

• **The filter must never**: invent facts, numbers, quotes, names, dates, or causal claims; change attribution; shift timeframes; change modality (might to did, alleged to proven); or reduce necessary public-safety warnings.

2.2 Scope and interfaces

Inputs (minimum): raw_title, raw_body, source_name, source_url, published_at. Optional but strongly recommended: raw_deck/subhead, captions, pull quotes, embeds, tables/charts (structured if available), and metadata (author, locale, section, updates).

Outputs: (A) Clean article: clean_title, clean_body, optional clean_summary; (B) Transparency package: flagged spans, type IDs + severity + confidence, before/after diffs, decisions/rationales, and red-line validation results.

2.3 Canonical taxonomy binding (hard IDs)

Every detection MUST map to one canonical type ID and may include secondary tags, but exactly one primary type per instance.

Field	Type	Notes
type_id	string	Example: B.2.1 (stable, sortable)
l1	string	Level 1 category
l2	string	Level 2 category
l3	string	Level 3 category
label	string	Human-friendly name

Recommended stable ID scheme: A Attention & Engagement, B Emotional & Affective, C Cognitive & Epistemic, D Linguistic & Framing, E Structural & Editorial, F Incentive & Meta.

2.4 Pipeline architecture (deterministic + explainable)

• Preprocess (lossless): preserve raw text; decode HTML; remove publisher cruft into a separate bin; preserve paragraph boundaries and quotes.

• Segment: title, deck, lede, body, lists, captions, pull quotes, embeds, tables/charts text.

• Detect: run detectors in parallel; produce span-level manipulation instances with type IDs.

• Score: compute confidence and severity (segment-aware).

• Decide: choose action (remove/replace/rewrite/annotate/preserve) using policy matrix and exemptions.

• Rewrite: apply bounded rewrite operations and templates; preserve facts, attribution, modality, scope.

• Validate: red-line validator checks semantic invariants; fallback to safer mode if needed.

• Emit: clean article + transparency package.

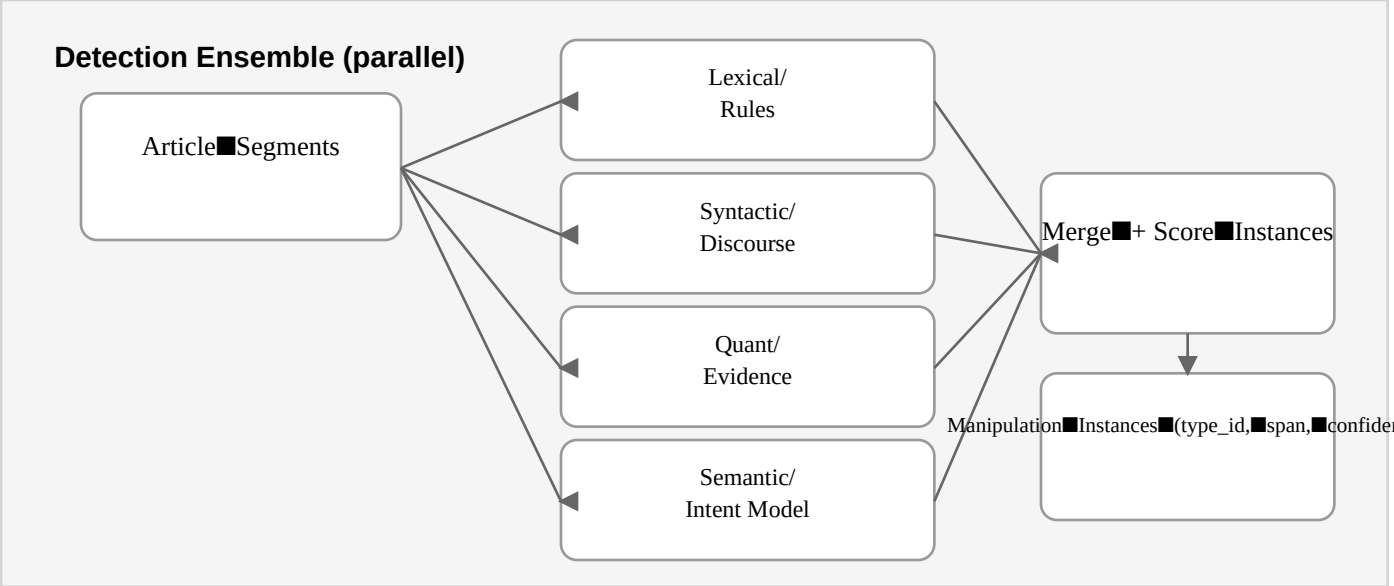

Figure 2. Parallel detection ensemble merged into scored manipulation instances.

2.5 Detection system (hybrid, layered)

Goal: high recall without destroying precision. Use a hybrid system with four detector classes; merge outputs into a single instance set.

• **(1) Lexical / pattern detector (rules)**: clickbait phrases, rage verbs, scarcity cues, vague authorities, weasel words, absolutes; punctuation inflation; ALL CAPS; open-loop patterns; quote-scoped exemptions.

• **(2) Syntactic / discourse detector**: passive voice + agent deletion; presuppositions and complex questions; rhetorical questions; narrative closure markers.

• **(3) Quant / evidence detector**: percent change without baseline; ‘surge/spike’ without magnitude; missing timeframe; correlation to causation language; single-study overreach; misleading averages (mean vs median); selection/survivorship cues; methodology opacity.

• **(4) Semantic / intent detector (model-assisted)**: subtle framing, minimization bias, false symmetry/proportionality, platform-native engagement bait, translation/interpretation bias signals, and insinuation detection.

Detection output contract (per instance):

Field	Type	Notes
detection_id	string	unique
type_id_primary	string	required, stable taxonomy ID
type_ids_secondary	string[]	optional
segment	enum	title/deck/lede/body/caption/pullquote/embed/table
span_start	int	character index in segment
span_end	int	exclusive
text	string	exact original span
confidence	float	0–1
severity	int	1–5
rationale	string	short, specific
recommended_action	enum	remove/replace/rewrite/annotate/preserve
rewrite_template_id	string	if rewrite/replace
exemptions_applied	string[]	e.g., inside_direct_quote

2.6 Scoring and decision policy

Confidence is the probability the span matches the type. **Severity** is the impact on reader autonomy and factual clarity.

• Severity 1: mild tone inflation, little meaning distortion.

• Severity 2: emotional push, still mostly factual.

• Severity 3: meaning-direction influence (framing materially shifts interpretation).

• Severity 4: epistemic distortion (certainty inflation, evidence laundering).

• Severity 5: manipulative core (pure engagement, dehumanization, conspiracy insinuation).

• Segment multipliers: Title x1.5; Deck x1.3; Lede x1.2; Captions x1.2; Body x1.0; Quotes x0.6 (but see quote rules).

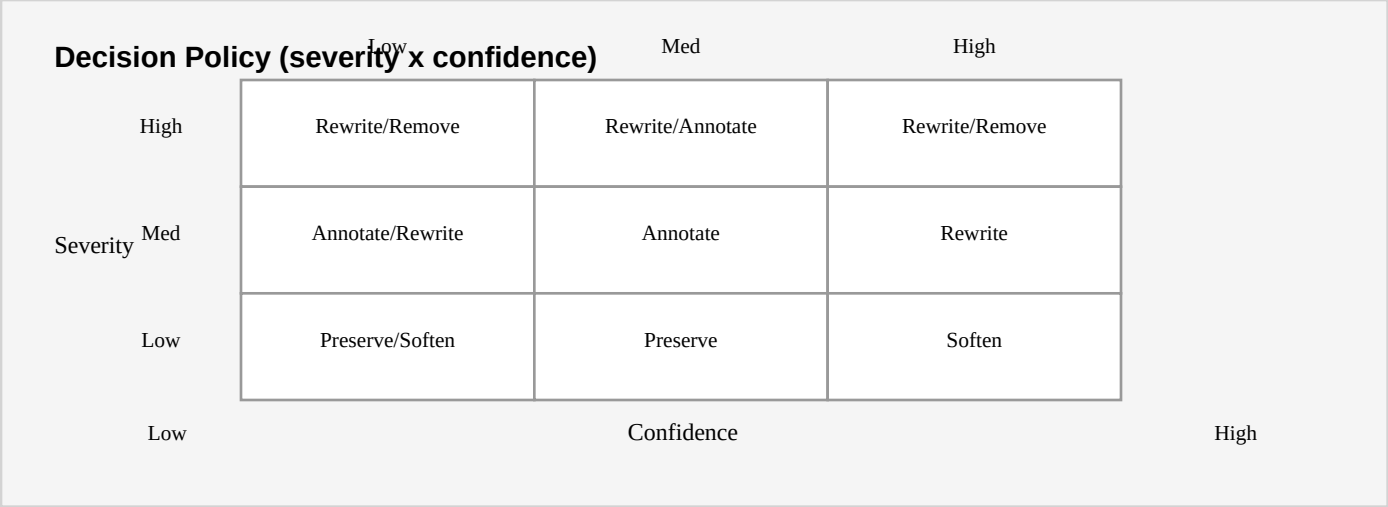

Figure 3. Decision policy heuristic. In implementation, enforce hard safety constraints before any rewrite.

Action rules:

• Severity 4–5 AND outside quotes → rewrite/remove (prefer rewrite if factual core exists).

• Severity 3 → rewrite or annotate.

• Severity 1–2 → replace/soften, or preserve if justified.

• Inside direct quote → preserve quote text; optionally annotate surrounding narration.

• Public-safety warnings → preserve meaning; trim flourish only.

Remove vs rewrite: Remove only when the span carries zero factual content (pure engagement). Rewrite when any factual content or legitimate uncertainty exists.

2.7 Rewrite engine (core IP behavior)

Rewrite principles:

• Factual invariance: keep entities, quantities, dates, and attributions unchanged.

• Modality preservation: alleged stays alleged; may stays may; ranges stay ranges.

• Attribution preservation: “X said/claimed/alleged” stays attributed.

• Scope preservation: geography, timeframe, population scope remain intact.

• Salience preservation: do not flatten legitimate stakes (harm, cost, safety).

Rewrite operations (atomic):

• Deintensify: remove hype adjectives/adverbs where they add no facts.

• Specify actor: replace vague authorities with named sources if present elsewhere; else annotate “not specified.”

• Replace rage verbs: “slams/blasts/destroys” → “criticized/said.”

• Remove open loops: insert the missing fact if in article; else state uncertainty without tease.

• Demetaphorize: convert war/contamination metaphors to literal description.

• Restore baselines: replace relative change with absolute baseline if present; otherwise flag baseline missing.

• Downgrade certainty: “proves” → “suggests/indicates,” consistent with evidence in text.

• De-binary framing: soften false dilemmas while preserving the underlying options described.

• Neutralize coded proxies (outside quotes): remove identity-coded persuasion where it's editorial, not factual.

• Extract claim from reaction: replace "outrage erupts" with "some users/officials criticized…" if evidence is present.

Hard rules (never do):

• Never introduce new facts or numbers.

• Never infer motives or intent.

• Never remove legally/materially relevant qualifiers ("alleged," "according to the complaint").

• Never rewrite a quote's meaning; quotes are preserved verbatim.

Quote handling (strict):

• Inside direct quotes: do not rewrite quote text.

• Allowed: remove sensational framing around the quote; add neutral attribution context; annotate as "quoted opinion" in transparency.

• Disallowed: changing meaning, tone, or implication inside the quote.

2.8 What NOT to remove (anti-sterility and anti-misleading guardrails)

A. Legitimate warnings

• Preserve clearly sourced public safety warnings, time-sensitive evacuation or health notices, and explicitly attributed risk assessments.

• Rule: If risk magnitude + source + timeframe are present (or official guidance is explicit), preserve; trim flourish only.

B. Genuine uncertainty

• Preserve "unknown," "unclear," "data incomplete," ongoing investigations, and credible conflicts in evidence.

• Rule: Uncertainty is information, not noise.

C. Moral language in quoted speech

• Preserve direct quotes expressing opinion; remove editorial endorsement around them.

D. Material conflict

• Preserve real disagreements and tradeoffs; remove inflation, not conflict.

E. Human stakes

• Preserve factual human impact and clearly labeled lived experience; remove emotional steering language.

F. Accountability and agency

• Do not blur actors or responsibility where the original is clear; remove only insinuation or hype.

G. Asymmetric reality

• Do not force balance when evidence is asymmetric; report confidence and sourcing.

2.9 Transparency layer (NTRL Report output)

Purpose: show readers and internal QA exactly what changed and why, without editorial judgment.

• Summary rollup: counts by L1 and by type_id; overall "manipulation density."

• Per-span table: original span, type_id + label, action taken, rewritten span, confidence + severity.

• Diff view: inline highlights (removed/replaced/annotated).

• Epistemic risk flags: anonymous-source heavy, missing baseline, headline mismatch, methodology not described, etc.

Field	Type	Notes
changes	array	list of change objects
summary_stats	object	counts by type_id, severity
redline_checks	object	pass/fail per check
version	string	filter version + taxonomy version
trace	object	optional decision trace for auditability

Change object schema:

Field	Type	Notes
detection_id	string	links back to detection
type_id_primary	string	primary taxonomy ID
before	string	exact span
after	string	rewrite result (or empty if removed)
action	enum	remove/replace/rewrite/annotate
reason	string	short, neutral explanation
location	object	segment + indices

2.10 Red-line validator (post-generation, must-pass constraints)

• Entity invariance: names/orgs/places unchanged unless correcting an obvious typo present in the source.

• Number invariance: numbers unchanged unless adding an explicit baseline present in the article.

• Date/time invariance: no shifting dates; no changing relative time claims unless the absolute date is present.

• Attribution invariance: who said/claimed/reported remains identical.

• Modality invariance: alleged/may/likely must not become did/will/confirmed.

• Causality invariance: no new causal links; preserve explicit caveats.

• Risk invariance: public safety risk not downplayed; official instructions preserved.

• Quote integrity: quoted text unchanged.

• Scope invariance: geography/timeframe/population preserved.

• Negation integrity: not/never/denies preserved.

Validator outputs: pass/fail; failed_checks[]; risk_level; recommended_fallback (annotate/preserve/partial_rewrite).

2.11 Structural and editorial manipulation handling

• Headline–body mismatch: rewrite headline to match body's certainty, modality, and timeframe; log in transparency.

• Omission bias within-article: if exculpatory or qualifying context exists later, ensure early summary does not omit it.

• Baseline surfacing: if a ‘surge/spike’ claim has small absolute numbers later, surface the absolute numbers earlier (without inventing).

• Quote/source choreography: detect quote ordering bias; in NTRL view, elevate factual context ahead of inflammatory quotes when appropriate, without removing quotes.

• Visual framing: if captions are present, remove emotive framing; keep factual description.

2.12 Publisher cruft handling

• Remove or separate: newsletter prompts, “read next,” autoplay prompts, shopping modules, polls, “watch next,” affiliate blocks.

• Log cruft removals in transparency as aggregated counts rather than overwhelming per-span noise.

2.13 Quality targets, telemetry, and evaluation

• Offline: faithfulness (no material change), calmness (reduced arousal terms), clarity (more explicit dates/actors/attribution when present), epistemic health (uncertainty preserved).

• Online: validator failure rates, high-diff alerts, user feedback (“too sterile” vs “still too spicy”), per-source anomaly detection.

• Feedback loop: human review UI for misclassified spans, rewrite improvements, and “should have preserved” flags; promote edits into rule/model/template updates.

2.14 Operational constraints and safe fallbacks

• Determinism and auditability: same input + same version → same output; prompts/version hashes stored; decision trace stored.

• Performance: fast path (rules + lightweight semantic) and slow path (deep checks for high-severity title/lede).

• Fallback triggers: extraction uncertainty high; model confidence low across many spans; repeated validator failures.

• Fallback actions: switch to Light NTRL mode; prefer annotate over rewrite; output minimal clean headline plus explicitly supported “What we know / What we don’t know” only when those items are grounded in the text.