SOP: Quality Control and Anti-Hallucination Protocol

*Generated: 03/09/2025 14:16*

# SOP: Quality Control and Anti-Hallucination Protocol

|  |  |
| --- | --- |
| **Document ID:** | **DWS-SOP-QUALITY-001** |
| \*\*Version:\*\* | 1.0 |
| \*\*Status:\*\* | Final |
| \*\*Approved By:\*\* | Craig Cottle |
| \*\*Date of Issue:\*\* | 26-Aug-2025 |
| \*\*Next Review Date:\*\* | 26-Feb-2026 |

---

## 1.0 Purpose

for all system outputs and forms the cornerstone of our quality assurance framework.

## 2.0 Scope

This SOP applies to all AI agents, orchestrators, human reviewers, and automated processes within the Autonomous Agentic Marketing System, including:

* All `SiteSpect`, `ContentForge`, and `StrategyNexus` squad outputs
* Data extraction from enhanced\_seo\_crawler.py and related crawling mechanisms
* Content generation, research synthesis, and strategic recommendations
* All intermediate processing steps and final deliverables
* Human-on-the-Loop review processes and approval gates

## 3.0 Definitions

* \*\*AI Hallucination:\*\* The generation of factually incorrect, fabricated, or unverifiable information by AI systems that appears plausible but lacks evidential foundation.
* \*\*Source Verification:\*\* The mandatory process of confirming all claims, statistics, and factual assertions against authoritative, traceable sources.
* \*\*Confidence Score:\*\* A numerical rating (0-100) indicating the system's certainty in data accuracy, with mandatory thresholds for different output categories.
* \*\*Quality Gate:\*\* A mandatory checkpoint where outputs must meet defined accuracy standards before progression to the next workflow stage.
* \*\*RAG Enhancement:\*\* Retrieval-Augmented Generation techniques that ground AI responses in verified source material.
* \*\*Multi-Agent Verification:\*\* Cross-validation of outputs by independent AI agents to identify inconsistencies and potential hallucinations.

## 4.0 Procedures

### 4.1 Procedure: Pre-Flight Data Validation Framework

All inputs to the system must undergo mandatory validation before processing begins.

### \*\*Step 1: File Format Validation\*\*

For all file-based inputs (CSV, PDF, XLSX):

1. \*\*Format Integrity Check:\*\*

* Verify file can be opened and parsed using designated libraries
* Confirm file structure matches expected schema
* Validate encoding (UTF-8 for text content)
* Log any parsing errors with specific error codes

1. \*\*Content Completeness Assessment:\*\*

* Check for missing required fields or columns
* Identify null values, empty cells, or malformed data
* Verify data types match expected formats
* Flag any suspicious data patterns that suggest corruption

1. \*\*Source Attribution Verification:\*\*

* Confirm file metadata includes source attribution
* Verify creation date and last modification timestamps
* Ensure file provenance is documented and traceable

### \*\*Step 2: URL and Web Data Validation\*\*

For all web-based inputs processed by enhanced\_seo\_crawler.py:

1. \*\*URL Accessibility Check:\*\*

* Verify URL returns 200 status code
* Confirm site is publicly accessible (not behind login/CAPTCHA)
* Check for robots.txt compliance
* Validate SSL certificate integrity

1. \*\*Content Extraction Validation:\*\*

* Implement fallback extraction strategies as defined in enhanced\_seo\_crawler.py
* Cross-verify critical data points using multiple extraction methods
* Flag any extraction failures or inconsistent results
* Apply confidence scoring to extracted data based on method reliability

### 4.2 Procedure: Multi-Layer Verification Protocol

Every AI-generated output must pass through multiple verification layers before approval.

### \*\*Layer 1: Source Grounding Verification\*\*

1. \*\*Primary Source Requirement:\*\*

* Every factual claim must be traced to a specific, authoritative source
* Source must be accessible and verifiable by human reviewers
* Publication date must be within relevance threshold (defined per content type)
* Source credibility must be established using E-E-A-T criteria

1. \*\*Citation Standards:\*\*

* Include full URL or complete bibliographic reference
* Specify exact page numbers or section references where applicable
* Note source access date for web-based references
* Flag any sources that cannot be independently verified

### \*\*Layer 2: Cross-Validation by Independent Agents\*\*

1. \*\*Independent Verification Process:\*\*

* Assign fact-checking to separate AI agent not involved in original generation
* Verification agent must use different data sources where possible
* Document any discrepancies between original and verification results
* Require resolution of all conflicts before proceeding

1. \*\*Consistency Analysis:\*\*

* Check internal consistency within single outputs
* Verify consistency across related outputs in the same workflow
* Flag contradictory statements or incompatible data points
* Require explanation for any identified inconsistencies

### \*\*Layer 3: Confidence Scoring and Threshold Management\*\*

1. \*\*Confidence Score Assignment (0-100 scale):\*\*

* \*\*90-100:\*\* Multiple authoritative sources, recent data, direct verification possible
* \*\*70-89:\*\* Single authoritative source, data reasonably current, indirect verification
* \*\*50-69:\*\* General source, older data, limited verification options
* \*\*Below 50:\*\* Insufficient sourcing, unverifiable claims, speculative content

1. \*\*Threshold Requirements by Output Type:\*\*

* \*\*Technical SEO Data:\*\* Minimum 85 confidence score
* \*\*Strategic Recommendations:\*\* Minimum 80 confidence score
* \*\*Content Substance:\*\* Minimum 75 confidence score
* \*\*General Research:\*\* Minimum 70 confidence score

1. \*\*Below-Threshold Handling:\*\*

* Outputs below minimum threshold automatically flagged for human review
* Must include specific identification of low-confidence elements
* Require additional sourcing or removal of unverifiable content
* Document rationale for any approved below-threshold content

### 4.3 Procedure: Anti-Hallucination Quality Gates

Mandatory checkpoints that prevent hallucinated content from advancing through the workflow.

### \*\*Quality Gate 1: Data Extraction Verification\*\*

Applied to all crawler and extraction outputs:

1. \*\*Extraction Method Cross-Check:\*\*

* Run identical extraction using at least two different methods
* Flag any discrepancies for manual review
* Require >95% consistency between methods for auto-approval
* Document extraction method reliability metrics

1. \*\*Plausibility Assessment:\*\*

* Apply statistical analysis to detect outlier data points
* Cross-reference against expected ranges for data type
* Flag unusual patterns that may indicate extraction errors
* Verify technical metrics against industry benchmarks

### \*\*Quality Gate 2: Content Generation Review\*\*

Applied to all AI-generated text content:

1. \*\*Factual Assertion Audit:\*\*

* Identify and catalogue all factual claims in generated content
* Verify each claim against documented sources
* Flag any claims lacking proper attribution
* Remove or request sourcing for unverifiable statements

1. \*\*Logical Consistency Check:\*\*

* Review argument flow and logical progression
* Identify contradictory statements within content
* Verify recommendations align with presented evidence
* Flag speculative content not clearly marked as opinion

### \*\*Quality Gate 3: British English Compliance Verification\*\*

Integrated quality control for language consistency:

1. \*\*Spelling and Grammar Standards:\*\*

* Apply British English spelling verification (colour, organisation, realise)
* Check punctuation conventions (single quotes for emphasis)
* Verify date formats (DD/MM/YYYY) and currency symbols (£)
* Ensure consistent terminology throughout outputs

1. \*\*Professional Communication Standards:\*\*

* Verify appropriate formality level for business communications
* Check adherence to DWS brand voice guidelines
* Ensure technical terminology consistency across outputs
* Validate abbreviation and acronym usage standards

### 4.4 Procedure: Human-on-the-Loop Quality Assurance

Mandatory human review checkpoints with defined escalation procedures.

### \*\*Tier 1: Automated Pre-Review\*\*

Conducted before human reviewer involvement:

1. \*\*Automated Quality Checklist:\*\*

* Verify all confidence scores meet minimum thresholds
* Confirm all factual claims include proper attribution
* Check British English compliance using automated tools
* Validate output completeness against defined requirements

1. \*\*Pre-Review Failure Handling:\*\*

* Outputs failing automated checks automatically returned to generation stage
* Specific failure reasons documented and provided to regeneration process
* Maximum three regeneration attempts before escalation to human review
* Track failure patterns to identify systemic issues

### \*\*Tier 2: Human Expert Review\*\*

Mandatory for all final outputs:

1. \*\*Review Assignment Protocol:\*\*

* Assign reviewers based on subject matter expertise
* Ensure reviewer independence from original generation process
* Provide standardised review checklist and quality criteria
* Set maximum review turnaround times (SLA): 4 hours for urgent, 24 hours for standard

1. \*\*Review Documentation Requirements:\*\*

* Document specific review actions taken
* Note any corrections or modifications made
* Assign final confidence rating to reviewed output
* Provide feedback to improve future generations

1. \*\*Review Escalation Process:\*\*

* Escalate to senior reviewer if initial review identifies major issues
* Require department head approval for any below-threshold approvals
* Document justification for any exception approvals
* Implement immediate process review for recurring quality failures

### 4.5 Procedure: Assumption Documentation Protocol

Mandatory documentation of all assumptions made during processing.

### \*\*Assumption Categories:\*\*

1. \*\*Data Assumptions:\*\*

* Document any missing data points and how gaps were addressed
* Note data quality limitations and potential impact
* Record any extrapolations or interpolations performed
* Specify data freshness assumptions and update requirements

1. \*\*Methodological Assumptions:\*\*

* Document analytical approaches and their limitations
* Note any simplifications made for computational efficiency
* Record alternative methods considered but not implemented
* Specify confidence levels for methodological choices

1. \*\*Context Assumptions:\*\*

* Document assumed audience knowledge level
* Note implied industry context or market conditions
* Record geographical or temporal scope assumptions
* Specify assumed business objectives or priorities

### \*\*Documentation Requirements:\*\*

* All assumptions must be explicitly stated in output metadata
* Provide rationale for each significant assumption
* Assess potential impact if assumption proves incorrect
* Include recommendations for assumption validation where possible

### 4.6 Procedure: Continuous Quality Monitoring

Ongoing assessment and improvement of anti-hallucination effectiveness.

### \*\*Quality Metrics Tracking:\*\*

1. \*\*Hallucination Detection Rate:\*\* Percentage of outputs flagged for factual inaccuracies
2. \*\*Source Verification Success:\*\* Percentage of claims successfully traced to authoritative sources
3. \*\*Cross-Validation Consistency:\*\* Agreement rate between independent verification attempts
4. \*\*Human Review Approval Rate:\*\* Percentage of outputs approved without modification
5. \*\*Processing Efficiency:\*\* Average time from input to final approved output

### \*\*Performance Review Schedule:\*\*

* \*\*Weekly:\*\* Review quality metrics and identify immediate issues
* \*\*Monthly:\*\* Assess threshold effectiveness and adjustment needs
* \*\*Quarterly:\*\* Comprehensive system performance evaluation and protocol updates
* \*\*Annually:\*\* Complete SOP review and update based on accumulated learnings

## 5.0 Integration Points

### 5.1 Enhanced SEO Crawler Integration

This SOP integrates with enhanced\_seo\_crawler.py anti-hallucination features:

* Utilises multi-strategy title extraction fallback mechanisms
* Applies confidence scoring to extraction method reliability
* Implements cross-validation of critical SEO metrics
* Documents extraction assumptions and method limitations

### 5.2 Technical Validation Test Suite Alignment

References and builds upon test\_seo\_validation.py protocols:

* Incorporates anti-hallucination test cases into quality gates
* Uses technical validation results to inform confidence scoring
* Applies error handling robustness tests to quality assessment
* Integrates validation testing into continuous monitoring framework

### 5.3 CLAUDE Test System Requirements

Aligns with CLAUDE\_TEST\_SYSTEM.md quality requirements:

* Implements structured testing protocols for quality verification
* Documents systematic approach to quality improvement
* Provides measurable quality metrics for system evaluation
* Ensures compatibility with existing technical testing infrastructure

## 6.0 Roles and Responsibilities

|  |  |
| --- | --- |
| **Role** | **Responsibilities** |
| \*\*System Orchestrator\*\* | Ensures all workflows implement mandatory quality gates |
| \*\*AI Agents\*\* | Apply confidence scoring and source verification to all outputs |
| \*\*Human Reviewers\*\* | Conduct expert review within defined SLA timeframes |
| \*\*Quality Assurance Lead\*\* | Monitor quality metrics and coordinate protocol improvements |
| \*\*Technical Lead\*\* | Maintain integration with technical validation systems |
| \*\*Project Manager\*\* | Ensure SOP compliance across all project deliverables |

## 7.0 Success Criteria

### 7.1 Zero-Tolerance Objectives

* \*\*Zero hallucinated data in final outputs:\*\* 100% of factual claims must be source-verified
* \*\*95%+ confidence scores:\*\* All approved outputs must meet minimum confidence thresholds
* \*\*100% source attribution:\*\* Every factual assertion must include verifiable source reference
* \*\*24-hour maximum review SLA:\*\* All outputs must complete quality review within defined timeframes

### 7.2 Operational Efficiency Targets

* \*\*<5% regeneration rate:\*\* Minimize outputs requiring multiple generation attempts
* \*\*>90% automated pre-review pass rate:\*\* Reduce manual review burden through effective automation
* \*\*<15% processing time overhead:\*\* Quality protocols should add minimal delay to overall workflow
* \*\*100% British English compliance:\*\* All outputs must meet linguistic and cultural standards

## 8.0 Risk Management

### 8.1 Critical Risks and Mitigation Strategies

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk** | **Impact** | **Probability** | **Mitigation Strategy** |
| \*\*Source Verification Failure\*\* | High | Medium | Mandatory dual-verification for critical claims |
| \*\*Review Bottleneck\*\* | Medium | High | Automated pre-screening and reviewer capacity planning |
| \*\*Technical System Failure\*\* | High | Low | Redundant validation systems and manual override protocols |
| \*\*Quality Threshold Conflicts\*\* | Medium | Medium | Clear escalation procedures and exception approval process |

### 8.2 Continuous Improvement Protocol

* Regular threshold calibration based on performance data
* Systematic analysis of quality failures to identify improvement opportunities
* Integration of new anti-hallucination techniques as they become available
* Stakeholder feedback integration for practical workflow improvements

---

* This SOP supersedes all previous quality control procedures
* Changes require approval from Quality Assurance Lead and Technical Lead
* All system users must acknowledge understanding of updated procedures
* Compliance monitoring is mandatory and subject to audit