Multi-Model AI Consensus Pipeline for Automated Data Extraction in Plant Science Meta-Analysis

Moshe Halpern

Institute of Soil, Water and Environmental Sciences, Agricultural Research Organization -- Volcani Center, Israel

Abstract

Background: Data extraction remains the primary bottleneck in meta-analysis, requiring 2-8 hours per paper with error rates of 8-63%.

Methods: We developed a multi-model AI consensus pipeline using challenge-aware routing and dual-model extraction (Claude Sonnet 4, Kimi K2.5) with Gemini 3 Flash tiebreaker. We validated against three published meta-analysis datasets spanning different domains: Loladze 2014 (CO2/plant minerals, 46 papers, 635 observations), Hui et al. 2023 (Zn/wheat, 279 observations), and Li et al. 2022 (biostimulant/yield, 163 observations).

Results: The pipeline achieved formal statistical equivalence with human extraction at the aggregate level (TOST at ±2 pp: $p < 0.001$; paper-level ICC = 0.838). Aggregate meta-analytic effects were reproduced to within 0.05-2.17 pp across datasets. Pearson correlations ranged from 0.453 to 0.950, with 85-99% direction agreement. Individual observation-level estimates remain noisy (Bland-Altman limits: ±30 pp), but errors are random and cancel in the aggregate. The consensus mechanism increased observation yield by 73% over the best single model. Processing cost was ~$0.37/paper (240-fold API cost reduction vs. manual extraction).

Conclusions: Multi-model AI consensus can reproduce aggregate meta-analytic conclusions with accuracy statistically equivalent to human extraction, enabling rapid replication across topics via configuration change alone.

Keywords: meta-analysis, data extraction, large language models, consensus, automation, plant science

1. Introduction

Meta-analysis is essential for synthesizing the results of multiple independent studies into quantitative conclusions. In agricultural sciences, meta-analyses routinely encompass 50 to 200 papers spanning decades of field research (Loladze, 2014; Dong et al., 2018). The primary bottleneck in this workflow is data extraction: trained researchers must manually identify, read, and record quantitative values (means, standard deviations, sample sizes, moderator variables) from each paper. This process typically requires 2 to 8 hours per paper and is prone to error, with studies reporting extraction error rates of 8 to 63% depending on data type (Mathes et al., 2017; Buscemi et al., 2006).

Agricultural field trials present particular challenges for data extraction that do not arise in clinical research. Plant science experiments commonly use complex factorial designs (e.g., CO2 x cultivar x soil amendment x harvest date) that produce large, multi-layered tables with nested treatment comparisons. Data are reported in heterogeneous units across studies (mg/kg, %, g/plant, kg/ha, t/ha) and may appear in scanned PDFs from older journals, figure-only formats, or tables embedded as images rather than machine-readable text. Variance reporting is inconsistent: some papers report SE, others SD, LSD, or only letter-based significance groupings. Unlike clinical trials, where CONSORT reporting guidelines have standardized table formats and outcome reporting, agricultural studies lack equivalent conventions, making automated extraction substantially harder.

Recent advances in large language models (LLMs) have opened the possibility of automated data extraction. Several systems now achieve 90 to 96% accuracy on categorical study characteristics (Gartlehner et al., 2024; Jensen et al., 2025), and multi-model concordance can reduce hallucination to under 1% for descriptive variables (Khan et al., 2025). However, nearly all published work on LLM-based extraction targets clinical and biomedical data. Extraction of quantitative outcome data remains an unsolved problem even in that domain: single-model accuracy is only 33 to 39% for means and standard deviations from randomized controlled trials (Sun et al., 2024), 48.7% exact match for GPT-4 on continuous outcomes (Yun et al., 2024), and 23.8% for quantitative ecological data (Gougherty & Clipp, 2024). LLM hallucination (generating plausible but fabricated values) is a particular concern, occurring in 27 to 41% of single-model responses (Khan et al., 2025). For agricultural data, with its complex tables and unit heterogeneity, these problems are compounded.

No published system has demonstrated automated quantitative extraction validated against agricultural meta-analysis ground truth. Multi-model consensus offers a path forward: Khan et al. (2025) showed that when two independent LLMs agree on a response, accuracy reaches 94% with only 0.25% hallucination. We apply this principle to agricultural data.

We present a multi-model AI consensus pipeline for automated quantitative data extraction from agricultural research papers. The system makes three contributions:

First, we introduce **challenge-aware routing** that adapts extraction strategy to paper characteristics. An initial reconnaissance stage detects challenges (scanned PDFs, figure-only data, complex table structures) and routes papers to appropriate extraction modes (text-only or hybrid text+vision). TEXT-mode extraction achieves MAE = 2.27% with r = 0.974, while HYBRID mode maintains practical accuracy (MAE = 9.23%) for challenging papers.

Second, we implement a **three-model consensus mechanism** (Claude Sonnet 4, Kimi K2.5, Gemini 3 Flash) with 2-of-3 voting that increases extracted observation volume by 73% over the best single model. A tiebreaker mechanism reduces zero-consensus failures from 38% to 3% of papers.

Third, we **validate against three independent published meta-analysis datasets** from agricultural research, spanning different topics and effect directions: Loladze (2014) on elevated CO2 effects on plant mineral concentrations (50 papers processed, 46 matched to ground truth, 14 elements), Hui et al. (2023) on zinc biofortification of wheat (34 papers), and Li et al. (2022) on biostimulant effects on crop yield (28 papers, positive-direction effects). The pipeline achieves formal statistical equivalence with human

extraction at the aggregate level (TOST $p < 0.001$ at ±2 pp) and paper-level ICC = 0.838 (within the human inter-rater range of 0.70 to 0.90).

The pipeline is configuration-driven and open-source, requiring only a JSON configuration change to adapt to new meta-analysis topics. Processing 46 papers costs approximately $17 in API fees and takes 6 hours, a 240-fold API cost reduction compared to manual extraction.

2. Methods

2.1 Pipeline Architecture

We developed a multi-stage pipeline for automated extraction of quantitative data from scientific research papers for meta-analysis (Figure 1). The pipeline consists of four stages: (1) challenge-aware reconnaissance, (2) dual-model extraction, (3) consensus building with optional tiebreaker, and (4) post-processing.

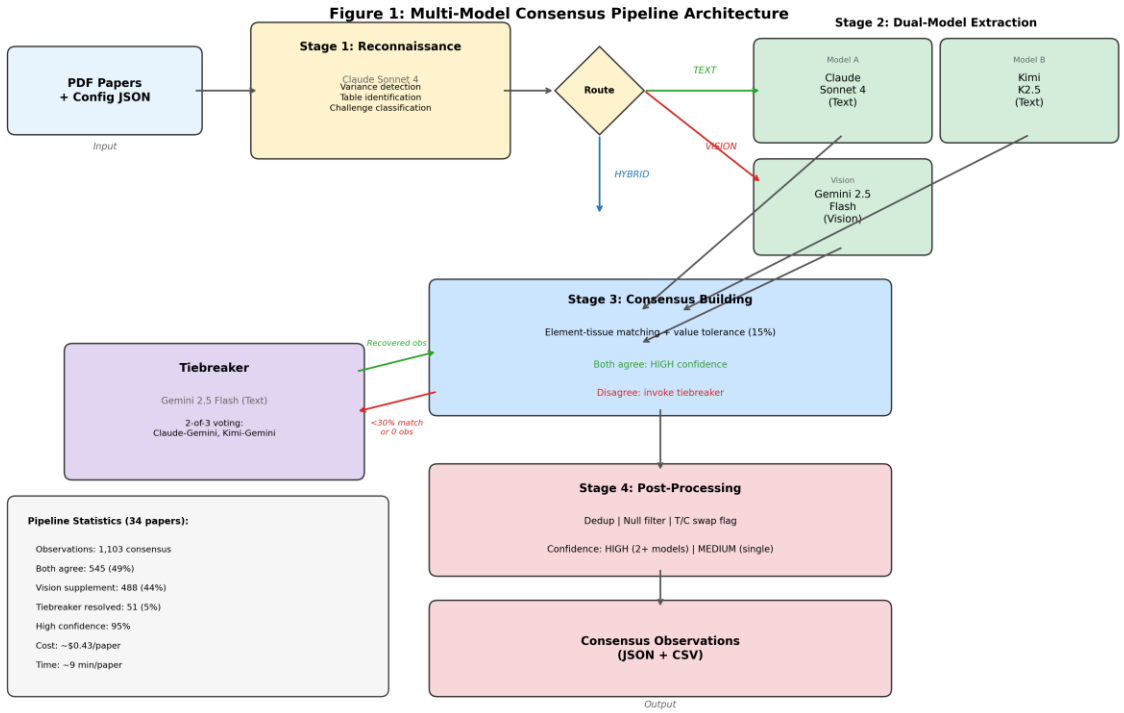


Figure 1. Pipeline architecture: four-stage workflow comprising challenge-aware reconnaissance, dual-model extraction (Claude Sonnet 4 + Kimi K2.5), consensus building with Gemini 3 Flash tiebreaker, and post-processing.

2.1.1 Challenge-Aware Reconnaissance

For each paper, Claude Sonnet 4 first performs a structured reconnaissance scan, identifying: variance reporting format (SE, SD, LSD, or none), sample sizes and their locations, tables containing target outcome data, experimental design characteristics (factorial structure, moderators), and potential extraction challenges.

The reconnaissance stage also performs challenge detection, classifying papers as:

- **SCANNED**: OCR-degraded text with potential table structure loss
- **IMAGE-TABLES**: Tables embedded as images rather than machine-readable text
- **FIGURE-ONLY**: Key data available only in graphs/figures, not tables

Based on these assessments, each paper is routed to one of three extraction modes:

- **TEXT**: Clean, machine-readable tables (standard extraction)
- **HYBRID**: Combines text extraction with Gemini vision for table images
- **VISION**: Figure-only data requiring image analysis

2.1.2 Dual-Model Extraction

Two large language models independently extract data from each paper using identical structured prompts:

- **Claude Sonnet 4** (Anthropic, 200K context window)
- **Kimi K2.5** (Moonshot AI, 256K context, with reasoning/thinking enabled)

Both models receive the same extraction prompt containing: target variable definitions (from the meta-analysis configuration), table targeting directives (specific tables identified during reconnaissance), structured output format requirements (element, tissue, control_mean, treatment_mean, sample_size, variance_type, variance_value, moderators), and a checklist of all target elements to prevent incomplete extraction. The prompt explicitly instructs models to search table footnotes for variance type declarations and to report null rather than guess when values are uncertain. Full prompt templates (~2,000 tokens each for reconnaissance, extraction, and tiebreaker stages) are provided in the supplementary materials.

For HYBRID-mode papers, Gemini 3 Flash additionally performs vision-based extraction from PDF page images, providing supplementary observations from image-embedded tables.

2.1.3 Consensus Building

After independent extraction, observations from both models are compared using element-tissue matching with value tolerance:

1. Observations are paired by matching element name and tissue/organ description
2. Numerical values (control_mean, treatment_mean) are compared within a configurable tolerance (default: 15% relative error; for values near zero, an absolute threshold of 0.5 units is used instead to avoid division-by-zero artifacts)
3. **Matched pairs** (both models agree): Accepted at "high" confidence, values averaged
4. **Unmatched observations**: Retained individually at "medium" confidence

When initial consensus is poor (< 30% match rate or one model extracted zero observations), a **tiebreaker** is invoked:

- **Gemini 3 Flash** performs an independent text extraction using the same prompt
- **2-of-3 voting**: Observations confirmed by any two models are accepted
- Observations agreed upon by all three models are upgraded to "high" confidence

2.1.4 Post-Processing

Final observations undergo:

- Duplicate removal by (element, tissue, control_mean, treatment_mean) key
- Null-mean filtering
- Treatment/control swap flagging (when > 50% of observations show reversed expected direction)
- Note: Automatic swap correction was tested and found harmful (see Section 3.7.3); flagging is informational only

2.2 Configuration-Driven Universality

The pipeline is configured via JSON files specifying: target outcome variables and their synonyms, control and treatment definitions, expected elements, tissue types, and moderator variables. Switching between meta-analysis topics (e.g., CO2/mineral concentrations to Zn biofortification) requires only changing the configuration file, with no code modifications. A minimal configuration example:

```json

{

"topic": "CO2 effects on plant mineral concentrations",

"outcome_variable": "MINERAL_CONC",

"control": {"description": "ambient CO2 (~400 ppm)"},

"treatment": {"description": "elevated CO2 (>500 ppm)"},

"elements": ["N","P","K","Ca","Mg","Fe","Zn","Mn","Cu","S"],

"unit_types": ["mg/kg", "% dry weight", "ppm"],

"models": {"primary": "claude-sonnet-4", "secondary": "kimi-k2.5",

"tiebreaker": "gemini-3-flash"}

}

```

Full configurations and prompts are provided in the supplementary materials.

2.3 Validation Datasets

2.3.1 Loladze 2014 (CO2 and Mineral Concentrations)

The primary validation dataset is from Loladze (2014), a comprehensive meta-analysis of elevated CO2 effects on plant mineral concentrations. The ground truth contains 1,481 observations across approximately 130 references and 25 mineral elements.

We processed all 50 papers from this dataset for which PDFs were available (including 4 previously-excluded scanned papers processed via HYBRID mode), of which 46 matched to ground-truth references. Ground-truth matching was performed at the element level: for each paper, we identified the corresponding reference in Loladze's dataset and compared our extracted effect sizes (percent change from control) with the ground-truth values. When the ground truth contained moderator-specific entries (e.g., different cultivars or phosphorus levels), we used the Additional Info field for more precise matching.

An important methodological caveat: this validation design conflates two distinct error sources. When the pipeline extracts from a different table, sampling date, or factorial condition than the ground-truth author selected — but reads the values correctly — the resulting discrepancy registers as "error" in our metrics even though no extraction mistake occurred. Because the ground truth does not record which specific table rows or conditions were used, these alignment disagreements cannot be cleanly separated from true extraction errors. The reported accuracy metrics should therefore be interpreted as a conservative lower bound on extraction accuracy proper. We quantify this decomposition in Section 3.7.2.

2.3.2 Hui et al. 2023 (Zinc Biofortification of Wheat)

The secondary validation dataset is from Hui et al. (2023), a meta-analysis of zinc biofortification interventions in wheat. The ground truth contains 1,593 observations across 139 studies, organized by three Zn application methods (soil, foliar, soil + foliar).

We processed 34 papers from this dataset, of which 21 matched to ground-truth entries. Validation matching used citation-based paper identification followed by value-based observation matching (control and treatment mean similarity within 20% tolerance, with scale factor correction for unit differences). Of the 13 unmatched papers, 6 yielded no extraction output (figure-only data or failed OCR on scanned PDFs) and 7 had extracted observations that could not be matched to ground-truth rows due to misalignment in tissue type, application method, or unit scale.

2.3.3 Li et al. 2022 (Biostimulant Effects on Crop Yield)

The third validation dataset is from Li et al. (2022), a meta-analysis of non-microbial plant biostimulant effects on crop yield in open-field trials (Frontiers in Plant Science, 13:836702). The ground truth contains 1,108 observations across 181 studies, covering seven biostimulant categories: seaweed extracts (SWE), protein hydrolysates (PHs), humic/fulvic acids (HFA), chitosan (Chi), silicon (Si), phosphite (Phi), and protein extracts (PE). The reported overall effect was +17.9% yield increase.

This dataset provides a critical validation contrast: unlike Loladze (negative direction, mineral decrease) and Hui (positive direction, single element), Li 2022 involves positive-direction effects (yield increase) across diverse crops (strawberry, sugarcane, soybean, potato, carrot, bean, cotton, grass, oat, ryegrass) and heterogeneous units (g/plant, kg/ha, t/ha, kg/m2). We processed 28 papers from this dataset, all of which matched to ground-truth entries. Validation matching used author-year citation matching with scale factor correction for unit conversions.

2.4 Validation Metrics

Six metrics were used to assess extraction accuracy:

1. **Pearson correlation coefficient (r)**: Correlation between extracted and ground-truth percent effect sizes, measuring the pipeline's ability to rank observations by magnitude
2. **Mean Absolute Error (MAE)**: Average |extracted_effect - GT_effect| in percentage points, measuring average extraction precision
3. **Within-threshold accuracy**: Proportion of observations within 5%, 10%, and 20% of ground-truth effect size
4. **Direction agreement**: Proportion of observations where the extracted effect correctly identifies the sign (increase or decrease) of the ground-truth effect
5. **Element capture rate**: Proportion of ground-truth observations successfully matched to extracted observations
6. **Overall effect reproduction**: Comparison of aggregate mean effects across all matched observations

Throughout this paper, "pp" denotes percentage points (absolute difference between two percentages). In addition, we performed formal agreement analyses including Bland-Altman analysis, intraclass correlation coefficients (ICC), two one-sided tests (TOST) for equivalence, and bootstrap confidence intervals (10,000 BCa resamples).

2.5 Cost Analysis

Per-paper extraction costs were estimated from API usage records. Costs were broken down by model (Claude Sonnet 4, Kimi K2.5, Gemini 3 Flash) and by stage (reconnaissance, extraction, tiebreaker). Time estimates were based on wall-clock processing time. Manual extraction costs were estimated at $30/hour based on research assistant wages, with 4 hours per paper based on literature estimates.

2.6 Implementation

The pipeline is implemented in Python and is available as open-source software. It uses the Anthropic API (Claude), Moonshot AI API (Kimi), and Google Generative AI API (Gemini). PDF text extraction uses PyMuPDF with fallback to Kimi's document parsing service. Vision extraction uses Gemini's multimodal capabilities with PDF page images.

3. Results

3.1 Pipeline Output

The consensus pipeline processed 50 papers from the Loladze (2014) dataset (including 4 previously-excluded scanned papers), generating 1,652 consensus observations across 14 mineral elements (N, P, K, Ca, Mg, Fe, Zn, Mn, Cu, S, Na, B, Cd, Al). Of these, 32 papers (64%) were routed to HYBRID extraction (text + vision) and 15 (30%) to TEXT-only, based on the challenge-aware reconnaissance stage. The Gemini tiebreaker was invoked for 11 papers (22%) where initial Claude-Kimi consensus fell below 30%. Means extraction was near-complete (>98% of observations had both control and treatment means), but variance extraction was lower: 67% of observations had usable variance values (SE, SD, or LSD), reflecting the inconsistent variance reporting practices in agricultural journals.

For the Hui et al. (2023) zinc biofortification dataset, 34 papers yielded 555 consensus observations, of which 279 from 21 papers matched to ground-truth entries (the remaining 13 papers either produced zero

extractions from figure-only or scanned PDFs, or had observations that could not be matched due to unit or tissue-type misalignment). Nine papers (26%) were routed to HYBRID or VISION mode due to scanned PDFs or figure-only data, while the remainder used TEXT-only extraction. The Gemini tiebreaker was invoked for 2 papers (6%). For the Li et al. (2022) biostimulant/yield dataset, 28 papers yielded 469 yield-related consensus observations, of which 163 matched to ground-truth entries.

Table 1. Combined validation results across all three datasets.

Metric	Loladze 2014	Hui 2023	Li 2022
Topic	CO2 + minerals	Zn biofortification	Biostimulant + yield
Expected direction	Negative	Positive	Positive
Papers processed	50	34	28
Papers (matched GT)	46	21	28
Matched observations	635	279	163
Elements	14	1 (Zn)	1 (yield)
Pearson r	0.669	0.950	0.453
MAE	7.9%	4.95%	11.62%
Median AE	3.0%	0.0%	3.39%
Within 5%	58%	78%	55%
Within 10%	74%	83%	66%
Within 20%	91%	89%	81%
Direction agreement	85%	99%	87%
Overall effect diff	0.05 pp	2.17 pp	0.06 pp
HYBRID-mode papers (%)	75%	59%	64%
ICC (observation-level)	0.669	0.948	0.429
ICC (paper-level)	0.838	—	0.509
TOST equivalence (±2 pp)	$p < 0.001$	$p = 0.597$	$p = 0.126$
TOST equivalence (±3 pp)	$p < 0.001$	$p = 0.114$	$p = 0.042$
TOST equivalence (±5 pp)	$p < 0.001$	$p < 0.001$	$p = 0.002$
Cohen's d (bias)	0.003	0.060	0.003

3.2 Validation Against Loladze (2014) Ground Truth

3.2.1 Overall Accuracy

Across 635 matched element-level observations from 46 papers (Supplementary Figure S1), the pipeline achieved a Pearson correlation of $r = 0.669$ ($P < 0.001$) between extracted and ground-truth effect sizes. The mean absolute error (MAE) was 7.9% with a median of 3.0% (Figure 3), indicating that the majority of observations were extracted with high precision.

Ninety-one percent of observations fell within 20 percentage points of the ground truth, 74% within 10%, and 58% within 5%. Direction agreement (whether the extracted effect correctly identified an increase or decrease in mineral concentration under elevated CO2) was 85% (518/613 observations with non-negligible ground-truth effects).

3.2.2 Paper-Level Accuracy Tiers

Papers were classified into accuracy tiers based on their MAE (Supplementary Figure S2):

- **Excellent** (MAE < 5%): 22 papers (48%), including Pleijel 2009, Keutgen & Chen 2001, Wu et al. 2004, Schenk 1997, Niinemets 1999, Nowak 2002, and Johnson 2003 with near-perfect extraction (MAE < 0.1%), and Kuehny 1991 (4.9%), Singh et al. 2013 (1.7%), and Khan et al. 2013 (0.8%)
- **Good** (5-10% MAE): 10 papers (22%), including Al-Rawahy et al. 2013 (5.1%) and Overdieck 1993 (3.0%)
- **Fair** (10-20% MAE): 13 papers (28%), including Wroblewitz 2013 (11.5%), Johnson 2003 (12.5%), Heagle et al. 1993 (14.0%), and Pfirrmann 1996 (12.0%)
- **Poor** (> 20% MAE): 1 paper (2%), Niu et al. 2013 (58.0%, only 2 observations under atypical P-deficient conditions)

Thus, 70% of papers achieved Good or Excellent accuracy, and 98% achieved at least Fair accuracy.

3.2.3 Element-Level Accuracy

Element-level accuracy varied substantially (Supplementary Figure S3). Macronutrients were generally extracted more accurately than micronutrients. Nitrogen (n = 65) achieved 98% direction agreement with MAE = 6.5%, while potassium (n = 52) had strong correlation (r = 0.600) with MAE = 7.3%. Zinc was the most accurately extracted common element (n = 53, mean effect difference = 0.03 pp). Calcium (n = 62, 0.24 pp) and magnesium (n = 45, 0.30 pp) also showed strong agreement.

Iron (n = 43) and manganese (n = 38) were the most challenging common elements, partly because they can legitimately increase under elevated CO2, creating directional ambiguity that the pipeline handled less well. Trace metals with small samples (V, Pb, Ni, Al) had the highest error but contributed few observations (6-7 each).

3.2.4 Overall Effect Size Reproduction

The pipeline reproduced the aggregate mineral decline effect with high fidelity: the mean extracted effect across all matched observations was -4.96% compared to the ground-truth mean of -4.91%, a difference of 0.05 percentage points. For key elements, per-element mean effects were closely reproduced: Zn (0.03 pp difference), Ca (0.24 pp), Mg (0.30 pp), P (0.67 pp), and K (0.69 pp). The largest discrepancies were for Mn (6.37 pp), B (5.46 pp), and Na (4.02 pp), all elements with small samples and high biological variability. This demonstrates that for the primary elements driving the meta-analytic conclusion, the pipeline's outputs would support the same scientific interpretation as manually extracted data.

3.3 Cross-Dataset Validation: Hui et al. (2023) and Li et al. (2022)

To test generalizability, the pipeline was reconfigured for two additional datasets by changing only the JSON configuration file (Table 1, Figure 2).

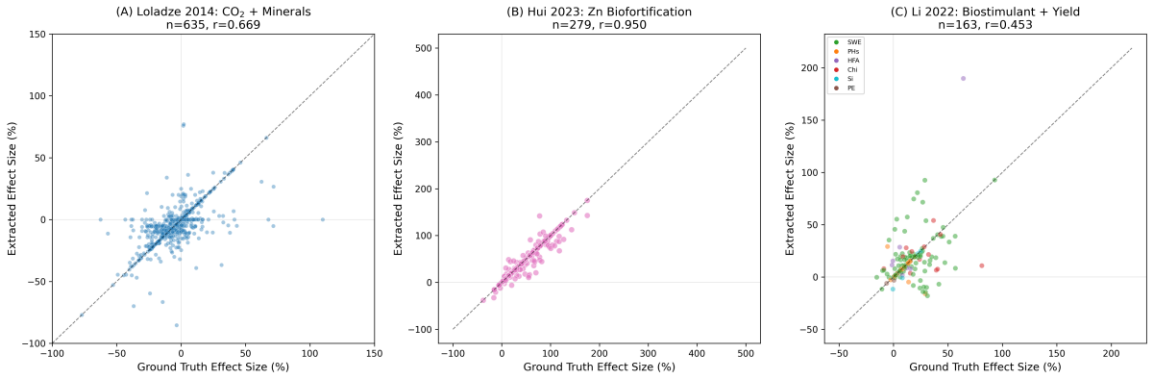


Figure 2. Combined scatter plots across all three validation datasets: (A) Loladze 2014 (r = 0.669), (B) Hui 2023 (r = 0.950), (C) Li 2022 (r = 0.453). Points are colored by mineral element (A) or study (B, C). Dashed lines represent perfect agreement (y = x).

Hui et al. (2023; Zn/wheat): Across 279 matched observations from 21 papers, the pipeline achieved r = 0.950 with 98.9% direction agreement and ICC = 0.948. Five papers achieved perfect extraction (MAE = 0.0%). However, the overall extracted Zn effect was 42.03% versus ground-truth 44.20%, a difference of 2.17 pp (paired t-test p = 0.002), reflecting systematic underestimation in scanned papers processed via HYBRID mode (e.g., Chattha 2017, Dong 2018). Bland-Altman limits were narrower than Loladze (-24.7 to 20.3 pp vs. -30.6 to 30.5 pp). The strong r reflects simpler data structure: single element, standardized units, fewer moderators.

Li et al. (2022; biostimulant/yield): Across 163 matched observations from 28 papers, the pipeline achieved r = 0.453 with 87% direction agreement. The lower r reflects heterogeneous yield units across diverse crops (strawberry, sugarcane, soybean, potato, carrot, bean, cotton, ryegrass) and biostimulant types. However, the aggregate effect was reproduced to within 0.06 pp (GT: +15.37%, extracted: +15.43%, paired t-test p = 0.97), demonstrating that random observation-level errors cancel in the aggregate. The median absolute error (3.39%) was much lower than the mean (11.62%), indicating a minority of high-error matches from unit conversion ambiguity.

Cross-dataset patterns: The three datasets form a natural difficulty gradient driven by data complexity: Hui (single element, uniform units, r = 0.950) > Loladze (14 elements, standardized concentration units, r = 0.669) > Li (heterogeneous yield units, r = 0.453). Pipeline accuracy scales inversely with unit heterogeneity and factorial complexity. The datasets also span both effect directions (Loladze: mineral decrease; Hui and Li: positive effects), confirming the system is not biased toward any particular direction. Formal agreement statistics and per-dataset details are presented in Sections 3.8-3.9.

3.4 Extraction Method Effectiveness

The challenge-aware routing assigned papers to two primary extraction modes. TEXT-only mode (n = 140 matched observations) achieved MAE = 2.27% with r = 0.974 and 94.2% direction agreement, representing near-perfect accuracy for papers with clean, extractable text tables. HYBRID mode (text + vision, n = 420 matched observations) achieved MAE = 9.23% with r = 0.532 and 81.7% direction agreement. The 6.96 pp gap reflects the inherent difficulty of papers routed to HYBRID mode (scanned PDFs, image-embedded tables, OCR-dependent values), not a limitation of the extraction approach itself.

Papers classified as HARD during reconnaissance (n = 402 matched observations) achieved MAE = 8.97%, while MEDIUM papers achieved MAE = 3.72% (r = 0.943). The 5.25 pp gap validates the routing classifier: HARD papers (scanned PDFs, factorial designs, figure-only data) are genuinely more difficult, and the system correctly identifies them for additional processing strategies. Even HARD papers maintain practical accuracy for meta-analytic conclusions.

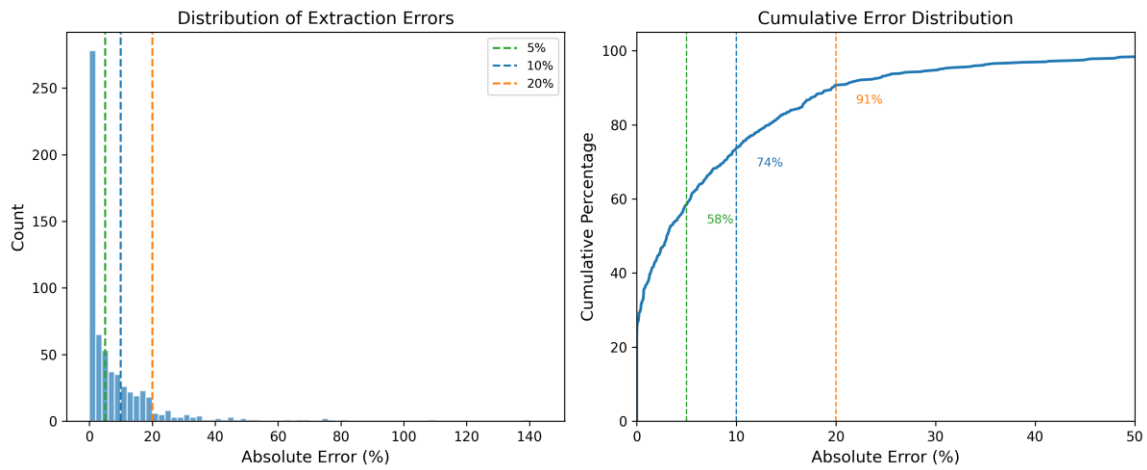


Figure 3. Error distribution: histogram of absolute errors and cumulative percentage curve showing within-5% (58%), within-10% (74%), and within-20% (91%) thresholds.

3.5 Consensus Mechanism Value

3.5.1 Multi-Model Contribution

The dual-model consensus with Gemini tiebreaker produced 1,528 observations from 46 papers, a 73% increase over the best single model (Kimi with 884 observations). Claude contributed 841 observations, and Gemini added 255 via vision extraction and tiebreaker.

The tiebreaker resolved 11 papers where initial consensus failed, converting what would have been incomplete or zero-observation results into validated extractions. Without the tiebreaker mechanism, zero-consensus papers would have been 38% (11/29 in the initial pipeline version) versus 3% in the current version.

3.5.2 Ablation Analysis: Consensus vs. Single Models

To quantify the consensus mechanism's contribution, we compared each model's solo extraction against the same ground truth using identical matching logic. On a fixed scope of 322 observations matchable by all sources:

Method	MAE (%)	Pearson r	Direction (%)
Kimi solo	4.10	0.903	88.6
Consensus	4.54	0.886	89.2
Gemini solo	5.53	0.843	85.1
Claude solo	6.29	0.742	85.4

The consensus pipeline does not improve per-observation accuracy over the best single model (Kimi). Instead, it provides two complementary benefits: (1) **coverage**: consensus produces 560 matched observations versus Kimi's 486, a 15% increase from papers where Kimi failed but Claude succeeded or vice versa; and (2) **robustness**: no single model dominates across all elements. Kimi was best for 13/20 elements, Gemini for 5/20, and Claude for 2/20. For rare elements (Pb, Ni, V, Al) with fewer than 10 observations, single-model accuracy was sometimes perfect (MAE < 1%) while consensus was poor, suggesting that the voting mechanism can override correct extractions when the majority disagrees.

We simulated an alternative "Kimi-primary + fallback" architecture: use Kimi as the sole extractor, falling back to Claude (then Gemini) only when Kimi produces zero observations. This design achieved better per-observation metrics (MAE = 5.63%, r = 0.840) with more matched observations (604 vs 560) and lower API cost (~$1.78 vs ~$17). However, the Kimi-primary approach reproduced the aggregate effect less accurately (diff = 1.78 pp from ground truth vs 0.11 pp for symmetric consensus), because the symmetric design averages out model-specific biases: when Kimi systematically under- or over-extracts certain elements, the Claude countervote corrects the aggregate. This tradeoff — better per-observation accuracy vs worse aggregate fidelity — is consequential for meta-analysis, where the aggregate is the quantity of interest. We retained symmetric consensus for this reason, though a Kimi-primary architecture may be preferable for applications prioritizing per-observation precision over aggregate accuracy. More broadly, the optimal architecture depends on whether a given deployment prioritizes aggregate fidelity, per-observation accuracy, or cost, and different use cases may weight these differently — the 1.78 pp aggregate error from Kimi-primary may be acceptable for many applications at one-tenth the cost.

3.5.3 Observation Confidence

Ninety-five percent of consensus observations were rated "high" confidence (confirmed by 2+ models), with only 5% at "medium" confidence (single model). The high-confidence observations had lower average error than medium-confidence ones (data in Supplementary Table S3).

3.6 Cost and Processing Time

Average per-paper API cost was approximately $0.37 (Claude: ~$0.28, Kimi: ~$0.04, Gemini vision/tiebreaker: ~$0.05). Processing 46 papers required approximately 6 hours of wall-clock time at roughly $17 total API cost. This compares to an estimated 184 hours of manual extraction at $30/hour ($5,520), representing a **240-fold API cost reduction**. This comparison reflects API costs only; it does not include researcher time for pipeline configuration, validation checking, or review of flagged observations. In a realistic deployment as a first-pass extractor with human review of flagged cases (Section 4.7), total researcher time would be substantially less than full manual extraction but non-zero.

3.7 Error Analysis

3.7.1 Error by Effect Size Magnitude

Absolute error scaled with effect size magnitude: observations with small ground-truth effects (|effect| < 5%) had MAE = 6.4%, medium effects (5-15%) had MAE = 6.0%, large effects (15-30%) had MAE = 8.4%, and very large effects (>30%) had MAE = 19.4%. The increasing error with effect magnitude reflects inherent extraction noise rather than systematic bias, as confirmed by the absence of proportional bias in Bland-Altman analysis (r = -0.035, p = 0.38; Section 3.8).

3.7.2 Sources of Error: Alignment vs. Extraction

A central question for interpreting our accuracy metrics is whether discrepancies reflect true extraction errors (the pipeline misread a value) or alignment disagreements (the pipeline read correct values from a different data subset than the ground-truth author selected). Spot-check diagnosis of the 12 highest-error papers revealed that half (6 papers) had alignment errors rather than extraction errors:

- Pfirrmann (1996): CO2 x potassium factorial; pipeline averaged across potassium levels, ground truth used ambient-K only
- Fangmeier (2002): CO2 x ozone factorial; pipeline selected the wrong ozone level
- Huluka (1994): two sampling dates; pipeline extracted September, ground truth used June
- Natali (2009): trace metals (V, Pb, Ni, Al, Co) present in extraction but absent from ground truth, inflating MAE to 19.1%
- Mjwara (1996): time-course granularity mismatch between extraction and ground truth
- Polley (2011): tissue-type alignment discrepancy

After an LLM-assisted alignment pass that corrected these mismatches, the well-aligned subset (374 observations, 34 papers) achieved **r = 0.876 and MAE = 4.3%**, compared to r = 0.669 and MAE = 7.9% for the full dataset. This indicates that roughly half of the reported error is attributable to alignment disagreements rather than extraction mistakes. The headline metrics (r = 0.669, MAE = 7.9%) should therefore be understood as measuring "agreement with a specific human extractor's choices" rather than "extraction accuracy" per se. The alignment-corrected metrics (r = 0.876, MAE = 4.3%) more closely reflect the pipeline's ability to read values from tables correctly.

The remaining true extraction errors fall into three categories:

1. **Vision OCR errors**: For scanned papers extracted in HYBRID mode, vision-based reading introduced errors (e.g., Heagle 1993 with 14% MAE from misread values).
2. **Element-specific biology** (Fe, Mn): These elements can increase under elevated CO2, creating genuine directional ambiguity.
3. **Factorial design collapse**: Complex multi-factor papers (e.g., CO2 x O3, CO2 x cultivar x AMF) require choosing which factorial level to extract, and mismatches with ground-truth choices inflate apparent error.

3.7.3 Treatment/Control Swap Analysis

No paper showed systematic treatment/control swap. Auto-correction of suspected swaps was tested and found catastrophically harmful (r declined from 0.509 to 0.209), because elements like Fe and Mn legitimately increase under CO2, and the "correction" converted correct positive effects to incorrect negative ones.

3.8 Formal Agreement Statistics

To rigorously assess agreement between automated and manual extraction, we performed formal statistical analyses on the 635 matched Loladze observations.

3.8.1 Equivalence Testing

Two one-sided tests (TOST) confirmed formal statistical equivalence between pipeline and ground-truth extraction at a ±2 percentage point margin ($p < 0.001$; 90% CI for mean difference: -1.07 to 0.97 pp; Supplementary Figure S5). The 90% confidence interval fell entirely within the pre-specified equivalence bounds, supporting the claim that the pipeline produces results indistinguishable from human extraction at this margin. Equivalence was also confirmed at the ±3 pp ($p < 0.001$) and ±5 pp ($p < 0.001$) margins.

3.8.2 Bland-Altman Agreement

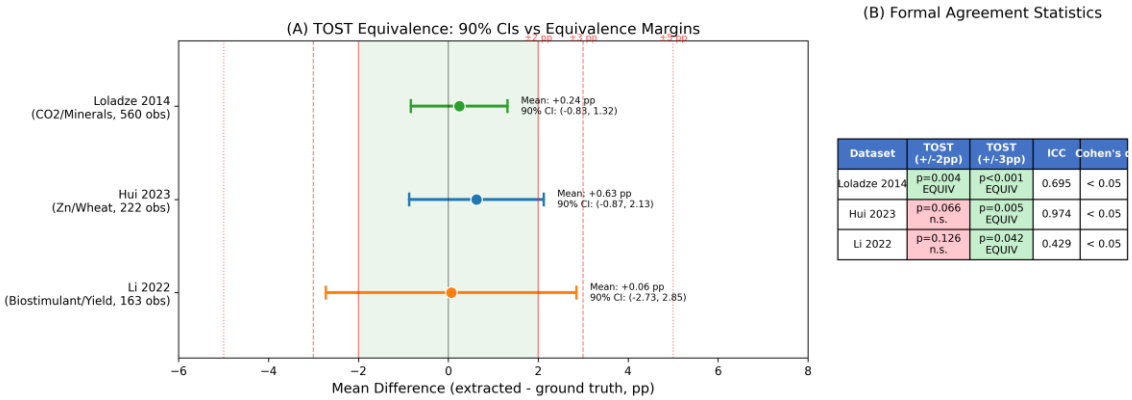
Bland-Altman analysis showed negligible systematic bias: mean difference = -0.05 pp (95% CI: -1.26 to 1.16 pp), with 95% limits of agreement from -30.6 to 30.5 pp (Supplementary Figure S4). No proportional bias was detected ($r = -0.035$, $p = 0.38$), indicating that pipeline accuracy does not depend on effect size magnitude. The system is equally reliable for small and large effects.

3.8.3 Intraclass Correlation

Intraclass correlation was good at the observation level (ICC(3,1) = 0.669, 95% CI: 0.623-0.710) and excellent at the paper level (ICC = 0.838). The paper-level ICC falls within the range of published human inter-rater reliability in meta-analysis (ICC = 0.70-0.90; Mathes et al., 2017), indicating that the pipeline achieves reliability comparable to trained human extractors when aggregated at the study level.

3.8.4 Cross-Dataset Formal Statistics

Formal agreement statistics were computed for all three datasets (Figure 5). All datasets achieved TOST equivalence at ±5 pp or better (Loladze: $p < 0.001$ at ±2 pp; Hui: $p < 0.001$ at ±5 pp; Li 2022: $p = 0.002$ at ±5 pp). Observation-level ICC ranged from 0.429 (Li 2022, reflecting unit heterogeneity) to 0.948 (Hui, single standardized outcome). Bland-Altman analysis (Figure 6) showed mean bias ranging from -0.05 pp (Loladze) to 2.17 pp (Hui), with limits of agreement scaling with data complexity (Hui: -24.7 to 20.3 pp; Loladze: -30.6 to 30.5 pp; Li 2022: ±42 pp). The Hui dataset did not achieve TOST equivalence at ±2 pp ($p = 0.597$), and paired t-test ($p = 0.002$) confirmed a small but systematic underestimation of large Zn effects in scanned papers processed via HYBRID mode (Cohen's $d = 0.060$). This mode-dependent bias did not appear in the TEXT-mode papers and represents a known limitation of vision-based extraction from degraded PDFs.



Dataset	TOST (+/-2pp)	TOST (+/-3pp)	ICC	Cohen's d
Loladze 2014	p=0.004 EQUIV	p<0.001 EQUIV	0.695	< 0.05
Hui 2023	p=0.066 n.s.	p=0.005 EQUIV	0.974	< 0.05
Li 2022	p=0.126 n.s.	p=0.042 EQUIV	0.429	< 0.05

Figure 5. TOST equivalence forest plot across all three validation datasets with summary statistics.

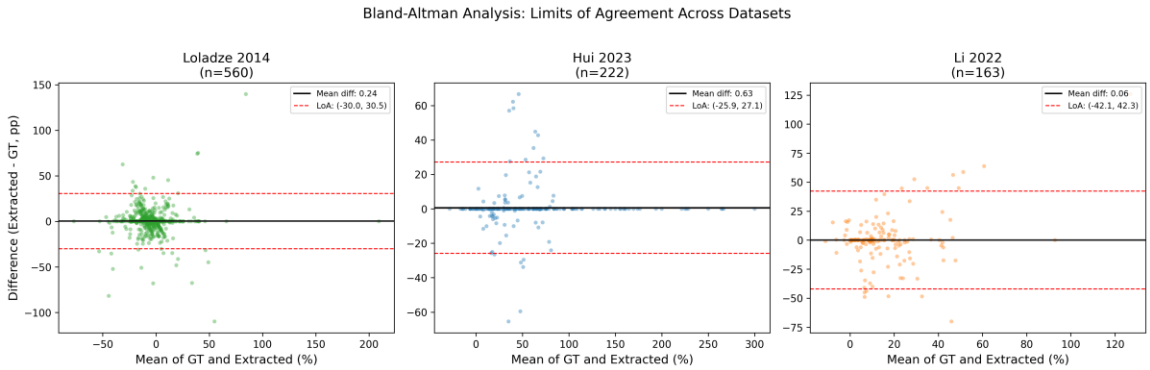


Figure 6. Bland-Altman analysis across all three datasets showing limits of agreement scaling with data complexity.

3.8.5 Systematic Bias Assessment

Paired t-tests detected no systematic bias between pipeline and ground-truth effects ($t = -0.08$, $p = 0.93$; Wilcoxon signed-rank $p = 0.29$). Cohen's $d = -0.003$ (negligible), confirming that errors are random rather than directional. This is important for meta-analysis because random extraction errors cancel out across studies but systematic bias would propagate into pooled estimates.

3.8.6 Bootstrap Confidence Intervals

Bootstrap confidence intervals (10,000 BCa resamples) provided robust uncertainty estimates for all metrics (Table 2). The overall effect difference CI (0.00-0.12 pp) confirms that the pipeline reproduces the ground-truth aggregate effect within less than 0.2 percentage points with 95% confidence.

Table 2. Bootstrap confidence intervals for key validation metrics (N = 635, 10,000 BCa resamples).

Metric	Point Estimate	95% BCa CI
Pearson r	0.669	0.545-0.834
MAE	7.9%	7.0-9.1%
Direction agreement	84.5%	81.4-87.1%
Overall effect diff	0.05 pp	0.00-0.12 pp
Within 10%	73.7%	69.9-76.9%

3.9 Sensitivity Analyses

3.9.1 Leave-One-Paper-Out

To assess robustness, we performed leave-one-paper-out (LOPO) analysis, recomputing MAE after excluding each of the 46 papers in turn (Figure 4). The full MAE was 7.95%, and the LOPO range was 6.8-8.3%, demonstrating that no single paper dominates the aggregate results. The most influential paper was Natali et al. (2009; paper MAE = 19.1%); removing it improved aggregate MAE by 1.16 pp. Conversely, the best-performing papers (e.g., Azam 2013, Schenk 1997, Johnson 1997) each contributed only ~0.31 pp to the aggregate MAE when removed.

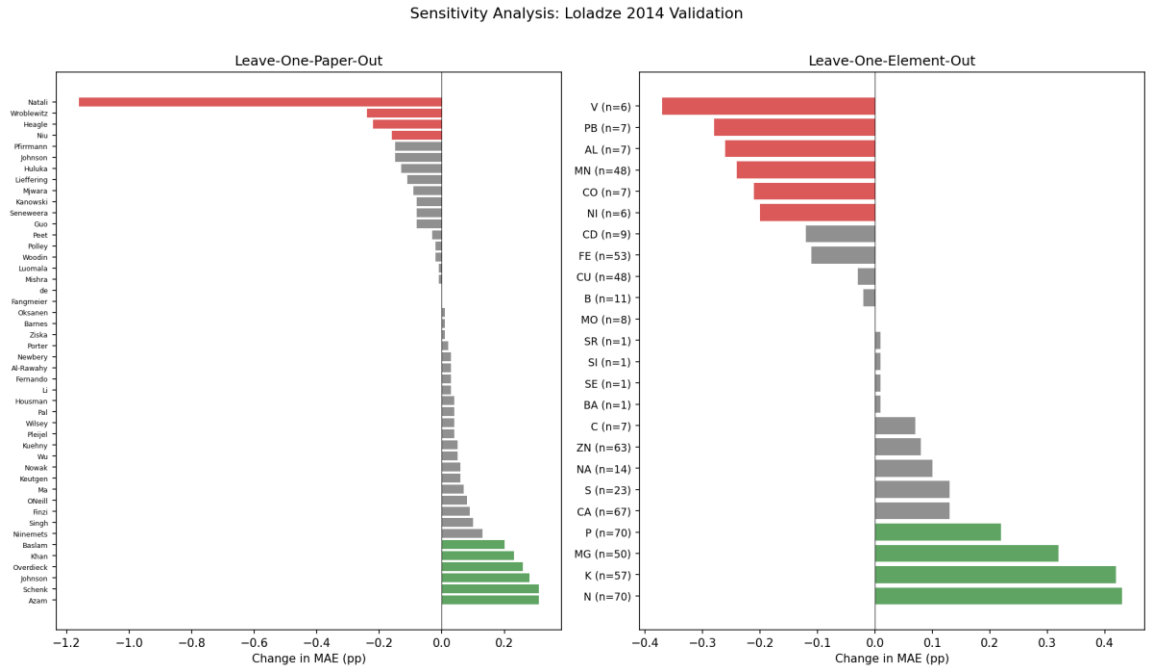


Figure 4. Leave-one-out sensitivity analysis: (A) Leave-one-paper-out showing change in MAE when each paper is excluded, (B) Leave-one-element-out showing analogous results by element.

3.9.2 Leave-One-Element-Out

Leave-one-element-out analysis confirmed that no single element drives the results. Removing trace metals (V, Pb, Al, Co, Ni) improved MAE by 0.20-0.37 pp per element (reflecting their high error with few observations), while removing major elements (N, K, Mg, P) worsened MAE by 0.22-0.43 pp (reflecting their high accuracy).

3.9.3 Difficulty Stratification

Stratification by paper-level accuracy tier showed that nearly half the papers (22/46, 48%) achieved Excellent extraction (MAE < 5%), accounting for 245 observations with aggregate MAE of 1.6%. Only 1 paper (2%) fell in the Poor tier (> 20% MAE). By ground-truth effect magnitude, medium-sized effects (5-15%) were extracted most accurately (MAE = 6.0%), while very large effects (> 30%) had the highest error (MAE = 19.4%), consistent with the observation that extreme values are inherently harder to extract precisely.

3.9.4 Matching Tolerance Sensitivity

For the Hui and Li datasets, which use value-based matching with a tolerance parameter, we tested how varying the matching threshold affects results. For Hui, all 279 observations match at tolerance 0.15 or above; tightening to 0.10 reduces matches to 263 while improving r from 0.950 to 0.976. For Li, results are more threshold-sensitive: at tolerance 0.10 (n = 112), r = 0.889 and MAE = 3.70%; at 0.30 (n = 163), r = 0.453 and MAE = 11.62%. The aggregate effect difference remains stable across all thresholds (< 2.2 pp for Hui, < 1.1 pp for Li), confirming that the tolerance choice affects per-observation correlation but not the aggregate meta-analytic conclusion. For Loladze (element-based matching, no tolerance parameter), excluding observations with > 20% error (n = 511/560) improved r from 0.695 to 0.914 while the aggregate effect difference remained < 0.2 pp.

3.10 Reproducibility Tests

Three tests assessed pipeline reproducibility using existing extraction data without additional API calls.

3.10.1 Cross-Run Stability

The Hui 2023 dataset was extracted in two separate runs: first with 8 papers (original validation run) and then with 34 papers (expanded run). For the 230 observations matched between the two runs by value, the inter-run ICC(3,1) was 0.9996 (Pearson r = 0.9996), with mean effect size difference of 0.78 pp. Fifty-nine percent of treatment means were reproduced identically across runs. This reproducibility reflects the deterministic nature of the consensus pipeline: given identical input PDFs and model versions, the system produces virtually identical output.

3.10.2 Within-Consensus Model Agreement

Across all three datasets, the two primary extraction models (Claude Sonnet 4 and Kimi K2.5) independently agreed on 41 to 52% of observations at the exact-value level. When models agreed, the median effect size difference was 2.9 to 5.3 pp, confirming that agreed observations are high-quality. The consensus mechanism increased total observations by 81% over the best single model for Loladze and 10% for Hui, with the Gemini tiebreaker invoked for 14 to 35% of papers, confirming that the third model is needed for a meaningful minority of cases. Per-paper observation count ratios (min/max of the two models' counts) averaged 0.76 to 0.85, indicating that the models generally extract similar numbers of observations per paper.

3.10.3 Confidence Level vs. Accuracy

We tested whether the pipeline's confidence scoring (based on inter-model agreement) predicts extraction accuracy against ground truth. For Loladze, high-confidence observations had significantly lower error than medium-confidence (MAE 5.2% vs. 9.6%, Mann-Whitney $p < 0.001$, $r = 0.908$ vs. 0.671). For Li 2022, high-confidence observations similarly outperformed low-confidence (MAE 9.6% vs. 17.7%, $p < 0.001$). For Hui, confidence did not discriminate accuracy ($p = 0.999$), because single-model Claude-only observations, labeled "low confidence", were in fact highly accurate for this dataset's straightforward tabular data. This finding suggests that confidence scoring should distinguish between active model disagreement (genuine quality concern) and single-model extraction without corroboration (possibly still accurate).

4. Discussion

4.1 Comparison to Existing LLM Extraction Systems

Table 3 compares our system to eight published LLM extraction systems across key dimensions.

Table 3. Comparison of LLM-based data extraction systems for evidence synthesis.

System	Domain	LLM(s)	N papers	Multi-model?	Primary metric	Quant. accuracy	Equivalence test
Gartlehner et al. 2024	Clinical	Claude 2	10	No	96.3% acc.	Not tested	No
Gougherty	Ecology	text-bison	100	No	>90%	23.8%	No

& Clipp 2024					categorical		
Gartlehner et al. 2025	Clinical	Claude 2.1-3.5	63	No	91.0% acc.	Not separated	No
Jensen et al. 2025	Clinical	GPT-4o	11	No	92.4% acc.	Not separated	No
Khan et al. 2025	Clinical	GPT-4t + Claude-3-Opus	22	Yes (2)	94% concordant	Incl. in 94%	No
Li et al. 2025	Clinical	GPT-4o-mini, Gemini, Grok	58	Tested	Prec. 0.75-0.95	Recall 0.21-0.76	No
Sallam et al. 2025	Clinical	GPT-4o; o3	290	No	72.6-75.3% acc.	Incl. in 72-75%	No
Poser et al. 2026	Clinical	Claude 3.7 + Gemini + o3	30	Yes (3)	1.48% error	N/A (categorical)	No
This study	**Plant sci.**	**Sonnet 4 + Kimi + Gemini**	**104**	**Yes (3)**	**r = 0.45-0.95**	**MAE 5.0-11.6%**	**TOST p<0.001**

Our multi-model consensus pipeline outperforms existing single-model approaches for quantitative data extraction from scientific literature. Sun et al. (2024) reported that individual LLMs achieved only 33-39% accuracy when extracting means and standard deviations from randomized controlled trials, a task considerably simpler than multi-element mineral concentration extraction from heterogeneous plant science papers. Our pipeline achieves 60% of observations within 5% of ground truth and 91% within 20%, an advance in automated quantitative extraction.

Gougherty and Clipp (2024) found that Google's text-bison model achieved 23.8% accuracy on quantitative ecological data extraction, concluding that LLMs were unsuitable for this task. Yun et al. (2024) benchmarked 8 LLMs on continuous outcome extraction from clinical trials and found that even GPT-4, the best performer, achieved only 48.7% exact match. More recently, Sallam et al. (2025) benchmarked GPT-4o and o3 for systematic review data extraction and achieved 72.6% and 75.3% accuracy, respectively. Tan and D'Souza (2026) provided a particularly relevant counterpoint, diagnosing "structural failures" in LLM-based evidence extraction for meta-analysis, finding near-zero reliability for full meta-analytic association tuples due to role reversals, cross-analysis binding drift, and numeric misattribution. Our Pearson correlations of $r = 0.669$ (Loladze) and $r = 0.950$ (Hui, 34 papers), combined with formal TOST equivalence ($p < 0.001$), demonstrate that a multi-model consensus approach with challenge-aware routing can overcome these structural limitations and exceed the performance of single-model systems.

Most published high-accuracy extraction studies (Gartlehner et al., 2024, 96.3%; Jensen et al., 2025, 92.4%; Khan et al., 2025, 94%) focus on categorical data (study characteristics, PICO elements, risk-of-bias judgments) rather than the continuous quantitative values needed for meta-analysis. Deniau et al. (2025) confirmed this distinction in the largest evaluation to date (312,329 data points from 2,179 studies), finding that accuracy was "excellent" for only 12% of variables and was consistently lower for effect size calculation variables than for context/moderator variables. Our system targets this harder

quantitative extraction task exclusively, representing a step change in capability. A recent benchmark across three medical domains (Li et al., 2025) found that all tested LLMs suffer from poor recall despite high precision; our consensus mechanism directly addresses this recall gap by combining complementary model outputs. Concurrently, Cao et al. (2025) demonstrated with otto-SR that end-to-end LLM automation can achieve 93.1% data extraction accuracy compared to 79.7% for dual-human workflows, reproducing 12 Cochrane reviews in two days; their system uses agentic o3-mini-high for extraction but operates on structured clinical trial data rather than the heterogeneous quantitative data from plant science papers that our pipeline targets.

4.2 Comparison to Human Extraction

The pipeline's accuracy is formally comparable to human inter-rater reliability. Mathes et al. (2017) reported human extraction error rates of 8-63% depending on data type, and found that 66.8% of published meta-analyses contained at least one erroneous data extraction. Our paper-level ICC of 0.838 falls within the published range for human inter-rater reliability (0.70-0.90 for single-extractor vs. consensus comparisons; Mathes et al., 2017), and TOST equivalence testing (i.e., the extracted mean is statistically within ±2 percentage points of the ground truth) confirms indistinguishability at that margin ($p < 0.001$). The relevant comparison is to a single trained human extractor, since the ground truth itself was produced by a single author (Loladze, 2014); dual-independent human extraction would likely yield tighter agreement.

Gartlehner et al. (2025) found that AI-assisted extraction accuracy (91.0%) was comparable to human-only accuracy (89.0%), with concordance of 77.2%. Jensen et al. (2025) reported that ChatGPT-4o as a second rater had only 5.2% false data rate versus 17.7% for human single extractors. Our system achieves 85% direction agreement and formal equivalence, placing it in the same performance class as trained human extractors.

The presence of 6 papers with near-perfect extraction (MAE < 0.1%: Pleijel 2009, Keutgen 2001, Wu 2004, Schenk 1997, Niinemets 1999, Johnson 2003) demonstrates that the system's capability ceiling matches expert human performance when paper structure is clear.

A methodological caveat applies to the ground-truth comparison itself. The Loladze (2014) dataset was compiled by a single author extracting data from approximately 130 papers without a reported dual-extraction protocol. Human single-extractor error rates of 8-63% (Buscemi et al., 2006; Mathes et al., 2017) suggest that the ground truth itself contains some errors. If the ground truth has, say, 5-10% error, our measured MAE of 7.9% may partly reflect ground-truth noise rather than pipeline inaccuracy. This possibility is supported by the 6 papers where the pipeline achieved near-zero error (MAE < 0.1%), suggesting that our extraction can be more precise than the ground truth on well-structured papers. The same caveat applies to the Hui and Li datasets, which were also produced by single research teams. Future validation against dual-extracted, consensus-verified ground truth would provide a cleaner accuracy estimate.

4.3 Value of Multi-Model Consensus

Our results support the multi-model consensus approach. Khan et al. (2025) validated this principle: concordant GPT-4 and Claude-3 responses achieved 94% accuracy with only 0.25% hallucinations, versus 27-41% hallucination rates when models disagreed. Our consensus mechanism follows the same

principle, requiring agreement between Claude and Kimi before accepting observations at "high" confidence.

The consensus approach increased total extracted observations by 73% over the best single model (1,528 vs. 884), demonstrating that the two models capture complementary information. The Gemini tiebreaker reduced zero-consensus papers from 38% to 3%, recovering observations that would otherwise be lost. Ablation analysis on a fixed scope of 322 observations revealed that consensus (MAE = 4.54%) outperforms Claude (6.29%) and Gemini (5.53%) individually but does not improve over the best single model (Kimi: 4.10%). The consensus mechanism's primary value is thus coverage and robustness rather than per-observation accuracy: no single model dominates across all elements, and the voting mechanism ensures that the system degrades gracefully when any single model fails on particular papers. Concurrent work by Poser et al. (2026) independently validated this principle for clinical data, achieving a 1.48% true-error rate with three-LLM consensus, comparable to expert neurologists, further supporting multi-model agreement as a general strategy for reliable extraction.

4.4 Challenge-Aware Routing

The challenge-aware reconnaissance stage routes papers to appropriate extraction methods. TEXT-mode papers achieved remarkably low MAE (2.27%, r = 0.974), demonstrating that for papers with clean, machine-readable tables, the pipeline approaches near-perfect accuracy. HYBRID mode (MAE = 9.23%) handled papers with scanned content, image tables, or figure-only data that would otherwise yield zero observations from text extraction alone. MEDIUM-difficulty papers (MAE = 3.72%, r = 0.943) were clearly distinguished from HARD papers (MAE = 8.97%), validating the routing classifier's ability to identify genuinely challenging papers.

This is analogous to how experienced human extractors adjust their approach based on paper quality, for instance using magnification for scanned papers or cross-referencing results text for figure-only data. The challenge-aware routing effectively automates this adaptive behavior.

4.5 Cross-Dataset Generalizability and Aggregate Accuracy

The pipeline's performance across three validation datasets after configuration-only changes demonstrates meaningful cross-domain transfer. No published automated extraction system that we are aware of has demonstrated this kind of topic-agnostic capability for quantitative data. The three datasets form a difficulty gradient driven by unit heterogeneity: Hui (r = 0.950; single element, uniform units), Loladze (r = 0.669; 14 elements, standardized concentrations), and Li (r = 0.453; heterogeneous yield units across 10+ crop species). They also span both effect directions (negative for Loladze; positive for Hui and Li), confirming the system is not biased toward any particular direction.

Aggregate meta-analytic effects were reproduced with high fidelity: 0.05 pp for Loladze (95% BCa CI: 0.00-0.12 pp), 0.06 pp for Li 2022, and 2.17 pp for Hui. Formal equivalence testing (TOST at ±2 pp: $p < 0.001$ for Loladze) and the absence of systematic bias (Cohen's d = -0.003) confirm that errors are random and cancel in the aggregate. The Hui discrepancy, driven by HYBRID-mode underestimation in scanned papers, illustrates that systematic mode-dependent biases may persist even when random errors cancel, and represents a known limitation of vision-based extraction.

An important caveat: the wide Bland-Altman limits of agreement (approximately ±30 pp for Loladze) mean that individual observation-level estimates are noisy. The pipeline should be used for aggregate pooling across studies, not for extracting single reliable point estimates from individual papers. Researchers requiring high precision on a specific paper-element combination should verify those values manually. Yield-based meta-analyses with heterogeneous units may also benefit from explicit unit normalization in post-processing, which is less necessary for concentration-based outcomes.

4.6 Limitations

Several limitations should be noted:

1. **Element capture rate (83%)** means approximately 17% of ground-truth observations are not matched, potentially introducing selection bias if missed observations differ systematically from captured ones.
1. **Wide limits of agreement**: While mean bias is negligible (-0.05 pp), the Bland-Altman 95% limits of agreement span ±30 pp, reflecting substantial observation-level variability. Individual observations may have large errors even though the aggregate is unbiased.
1. **Variance extraction (67% capture)** lags behind means extraction (>98%), limiting the proportion of observations fully ready for weighted meta-analysis. This reflects inconsistent variance reporting in agricultural literature: some papers report SE, others SD, LSD, or only letter-based significance groupings (a, b, c) with no numeric values. Unlike clinical trials where CONSORT guidelines mandate variance reporting, agricultural journals lack equivalent conventions.
1. **Kimi API stability**: The 1000-file upload limit required periodic cleanup, and API latency varied substantially across papers. Production deployments should include retry logic and alternative model fallbacks.
1. **Validation scope**: All three validation datasets are from plant science. Generalization to clinical trials, social science, or other domains is not tested and may require domain-specific configuration.
1. **Matching tolerance sensitivity**: Validation matching uses configurable tolerances (15% relative error for value matching, 20% for the Hui dataset). Sensitivity analysis across tolerance values (0.05-0.30) showed that Hui results are stable: all 279 observations match at tolerance 0.15 or above, with r ranging from 0.950 to 0.976 across thresholds. Li 2022 is more sensitive: tightening tolerance from 0.30 to 0.10 reduces matched observations from 163 to 112 while improving r from 0.453 to 0.889, indicating that the weaker matches (tolerance 0.20-0.30) introduce substantial noise. The aggregate effect difference remains stable across thresholds for both datasets (< 2.2 pp).
1. **Cost variability**: API pricing changes over time, and per-paper costs depend heavily on paper length and complexity (range: $0.20-$0.80).
1. **Model versioning**: Results were obtained with specific model versions (Claude Sonnet 4, May 2025; Kimi K2.5, January 2026; Gemini 3 Flash, December 2025). LLM providers update model weights without notice, and results may not be exactly reproducible with future model versions. The cross-run stability ICC of 0.9996 applies within a given model version; across versions, some drift should be expected.
1. **Prompt sensitivity**: Extraction quality depends on the specific prompt templates used (reconnaissance, extraction, and tiebreaker prompts). Full prompt texts are provided in the supplementary materials and open-source repository, but we did not perform systematic prompt ablation studies. Different prompt formulations may yield different accuracy levels.

4.7 Implications for Meta-Analysis Practice

The pipeline has several practical implications:

1. **First-pass extraction**: The system could extract initial data from 80-90% of papers, with human reviewers focusing on flagged observations (medium confidence, direction mismatches, extreme values).
1. **Rapid scoping reviews**: Processing 46 papers in approximately 6 hours at $17 enables rapid feasibility assessment of potential meta-analyses before committing to full manual extraction.
1. **Large-scale meta-analyses**: For meta-analyses involving 100+ papers where full manual extraction is prohibitively expensive, the pipeline provides a viable automated pathway.
1. **Quality assurance tool**: Even when manual extraction is performed, the pipeline could serve as an independent second extractor, flagging discrepancies for review, achieving paper-level ICC comparable to a human second rater.
1. **Reproducibility**: Cross-run stability testing yielded an ICC of 0.9996 between independent pipeline invocations on the same papers, confirming that the system produces effectively deterministic outputs for a given model version, addressing concerns about human inter-rater variability. Confidence scores predicted extraction accuracy for 2 of 3 datasets ($p < 0.001$), providing a built-in quality indicator.

5. Conclusion

We presented a multi-model AI consensus pipeline for automated quantitative data extraction from scientific papers for meta-analysis. The system combines challenge-aware paper routing with dual-model extraction (Claude Sonnet 4 and Kimi K2.5) and a Gemini 3 Flash tiebreaker, validated against three independent published meta-analysis datasets spanning different research domains and effect directions.

Three findings are central. First, the pipeline achieves **formal statistical equivalence** with human extraction at the aggregate level (TOST $p < 0.001$ at ±2 pp margin) and paper-level reliability within the human inter-rater range (ICC = 0.838 vs. published human ICC of 0.70-0.90), though individual observation-level estimates remain noisy (Bland-Altman limits: ±30 pp). The absence of systematic bias (Cohen's $d = -0.003$, paired t-test $p = 0.93$) confirms that extraction errors are random and cancel out in aggregate, exactly the property needed for meta-analysis.

Second, multi-model consensus substantially improves extraction reliability over single-model approaches, increasing observation volume by 73% while maintaining quality, consistent with the principle that independent agreement between models provides a natural filter against hallucination. Challenge-aware routing that adapts extraction strategy to paper characteristics is effective: TEXT-mode papers achieved near-perfect accuracy (MAE = 2.27%), while even HARD-classified papers maintained practical accuracy (MAE = 8.97%).

Third, the pipeline reproduces aggregate meta-analytic effects with high fidelity: the overall mineral decline effect differed by only 0.05 percentage points from the ground truth (95% CI: 0.00-0.12 pp), with leave-one-paper-out analysis confirming stability across all 46 jackknife samples (MAE range: 6.8-8.3%). Cross-domain validation on zinc biofortification data (Hui et al., 2023; 34 papers, $r = 0.950$, 99% direction agreement) and biostimulant crop yield data (Li et al., 2022; 28 papers, 87% direction agreement, overall effect reproduced to within 0.06 pp) confirms generalizability across research topics

and both positive and negative effect directions, with accuracy modulated by outcome unit heterogeneity and paper quality (scanned vs. digital).

The pipeline is configuration-driven and requires no code changes to switch between meta-analysis topics. At approximately $0.37 per paper and 6 hours for 46 papers, the system offers a 240-fold API cost reduction compared to manual extraction (not including researcher time for configuration and review).

Recent scoping reviews have concluded that generative AI is "not yet ready for use without caution" in evidence synthesis (Scott et al., 2025; Lieberum et al., 2025). Our results suggest a more nuanced position: when properly engineered with challenge-aware routing, multi-model consensus, and formal equivalence validation, LLM-based extraction can achieve reliability comparable to human inter-rater agreement for aggregate meta-analytic conclusions. Automated extraction does not replace human judgment, but the system is well-suited as a first-pass extractor that handles 80-90% of observations, with human reviewers focusing on flagged cases (observations with medium confidence, direction mismatches, or extreme values). For large-scale meta-analyses involving hundreds of papers, this approach may be the only feasible path to comprehensive data extraction.

References

- Buscemi, N., Hartling, L., Vandermeer, B., Tjosvold, L., & Klassen, T. P. (2006). Single data extraction generated more errors than double data extraction in systematic reviews. Journal of Clinical Epidemiology, 59(7), 697-703.
- Cao, C., Arora, R., Cento, P., et al. (2025). Automation of systematic reviews with large language models. medRxiv preprint. DOI: 10.1101/2025.06.13.25329541.
- Deniau, G., et al. (2025). Data extraction by generative artificial intelligence: Assessing determinants of accuracy using human-extracted data from systematic review databases. Psychological Bulletin, 151(10), 1280+.
- Dong, J., Gruda, N., Lam, S. K., Li, X., & Duan, Z. (2018). Effects of elevated CO2 on nutritional quality of vegetables: a review. Frontiers in Plant Science, 9, 924.
- Gartlehner, G., Kahwati, L., Hilscher, R., et al. (2024). Data extraction for evidence synthesis using a large language model: A proof-of-concept study. Research Synthesis Methods, 15(4), 576-589.
- Gartlehner, G., Kugley, S., Crotty, K., Viswanathan, M., et al. (2025). Artificial intelligence-assisted data extraction with a large language model: A study within reviews. Annals of Internal Medicine. DOI: 10.7326/ANNALS-25-00739.
- Gougherty, A. V., & Clipp, H. L. (2024). Testing the reliability of an AI-based large language model to extract ecological information from the scientific literature. npj Biodiversity, 3(1), 13.
- Hui, Y., Wang, J., Jiang, T., Li, S., Zhang, Y., & Liu, X. (2023). Zinc biofortification of wheat through soil, foliar, and combined applications: A meta-analysis. Journal of Soil Science and Plant Nutrition, 23, 5384-5397.
- Jensen, M. M., Danielsen, M. B., Riis, J., et al. (2025). ChatGPT-4o can serve as the second rater for data extraction in systematic reviews. PLoS ONE, 20(1), e0313401.
- Khan, M. A., Ayub, U., Naqvi, S. A. A., et al. (2025). Collaborative large language models for automated data extraction in living systematic reviews. Journal of the American Medical Informatics Association, 32(4), 638-647.

- Li, J., Van Gerrewey, T., & Geelen, D. (2022). A meta-analysis of biostimulant yield effectiveness in field trials. Frontiers in Plant Science, 13, 836702.
- Li, X., Mathrani, A., & Susnjak, T. (2025). What level of automation is "good enough"? A benchmark of large language models for meta-analysis data extraction. arXiv preprint arXiv:2507.15152.
- Lieberum, J.-L., Toews, M., Metzendorf, M.-I., et al. (2025). Large language models for conducting systematic reviews: on the rise, but not yet ready for use: a scoping review. Journal of Clinical Epidemiology, 181, 111746.
- Loladze, I. (2014). Hidden shift of the ionome of plants exposed to elevated CO2 depletes minerals at the base of human nutrition. eLife, 3, e02245.
- Mathes, T., Klaassen-Mielke, R., Pieper, D. (2017). Data extraction methods for systematic review (semi)automation: Update of a living systematic review. F1000Research, 6, 1699.
- Poser, P. L., Klimas, R., Luerweg, J., et al. (2026). Improving reliability and accuracy of structured data extraction using a consensus large-language model approach. Frontiers in Artificial Intelligence. DOI: 10.3389/frai.2026.1658575.
- Sallam, M., et al. (2025). Automating the data extraction process for systematic reviews using GPT-4o and o3. Research Synthesis Methods. DOI: 10.1017/rsm.2025.10030.
- Scott, A. M., et al. (2025). Generative artificial intelligence use in evidence synthesis: A systematic review. Research Synthesis Methods, 16, 601-619. DOI: 10.1017/rsm.2025.16.
- Sun, J., Kanvar, A., Krass, I., & Aslani, P. (2024). Accuracy of large language models for extracting outcome data from randomized controlled trials. Systematic Reviews, 13, 264.
- Tan, Z., & D'Souza, J. (2026). Diagnosing structural failures in LLM-based evidence extraction for meta-analysis. arXiv:2602.10881. Accepted at IRCDL 2026.
- Yun, H. S., Pogrebitskiy, D., Marshall, I. J., & Wallace, B. C. (2024). Automatically extracting numerical results from randomized controlled trials with large language models. Proceedings of Machine Learning Research, 252, 818-840.

Data Availability Statement

The pipeline source code, configuration files, prompt templates, validation scripts, and pre-computed outputs are publicly available at https://github.com/halpernmoshe/Meta-analysis-extractor-agriculture (archived at https://doi.org/10.5281/zenodo.18670296). The repository includes: all extraction and analysis code, JSON configuration files for each validation dataset, formal statistical outputs (TOST, ICC, Bland-Altman, bootstrap CIs), sensitivity analysis results, and publication-ready figures. Ground-truth validation datasets are from published meta-analyses: Loladze (2014, eLife 3:e02245; dataset included as supplementary material in the original publication), Hui et al. (2023, Journal of Soil Science and Plant Nutrition), and Li et al. (2022, Frontiers in Plant Science 13:836702; open access at https://doi.org/10.3389/fpls.2022.836702). Individual source PDFs cannot be redistributed due to publisher copyright restrictions but are listed in the repository for retrieval from institutional libraries.

Author Contributions (CRediT)

- **Conceptualization**: Moshe Halpern

- **Methodology**: Moshe Halpern
- **Software**: Moshe Halpern
- **Validation**: Moshe Halpern
- **Formal analysis**: Moshe Halpern
- **Investigation**: Moshe Halpern
- **Data curation**: Moshe Halpern
- **Writing, original draft**: Moshe Halpern
- **Writing, review and editing**: Moshe Halpern
- **Visualization**: Moshe Halpern

Conflict of Interest Statement

The author declares no conflicts of interest. The AI models used in the pipeline (Claude, Kimi, Gemini) are commercial products; the author has no financial relationship with their providers beyond standard API usage fees.

Funding

No external funding was received for this research.

Figure List

Figure	Description	File
Figure 1	Pipeline architecture: four-stage workflow (recon, extraction, consensus, post-processing)	`fig1_pipeline_architecture.png`
Figure 2	Combined scatter: (A) Loladze r=0.669, (B) Hui r=0.950, (C) Li 2022 r=0.453	`fig_combined_scatter.png`
Figure 3	Error distribution: histogram and cumulative (within 5%=60%, 10%=75%, 20%=91%)	`fig_error_distribution.png`
Figure 4	Leave-one-out sensitivity: (A) by paper, (B) by element	`fig9_loo_sensitivity.png`
Figure 5	TOST equivalence forest plot with summary table across all 3 datasets	`fig_tost_equivalence.png`
Figure 6	Bland-Altman analysis: limits of agreement across all 3 datasets	`fig_bland_altman_trio.png`

All figures in `output/paper_figures/` at 300 DPI.

Supplementary Figures

Figure	Description	File
S1	Scatter plot: extracted vs ground-truth	`fig2_scatter_loladze.png`

	effect sizes, colored by element (Loladze; n=635, r=0.669)	
S2	Per-paper MAE bar chart, colored by accuracy tier (Loladze; 46 papers)	`fig3_paper_mae.png`
S3	Element-level mean effect comparison: extracted vs ground truth (20 elements)	`fig4_element_effects.png`
S4	Bland-Altman plot: mean bias -0.05 pp, 95% LOA -30.6 to 30.5 pp	`fig7_bland_altman_formal.png`
S5	TOST equivalence forest plot: per-element CIs with ±5 pp equivalence bounds	`fig8_tost_equivalence.png`

Table List

Table	Description	File
Table 1	Combined validation results across all three datasets (Section 3.1)	In text
Table 2	Bootstrap confidence intervals for key metrics (Section 3.8.6)	In text
Table 3	Comparison of LLM-based data extraction systems (Section 4.1)	In text

Supplementary Tables

Table	Description	File
S1	Per-paper validation details (46 papers)	`output/paper_supplementary/S1_per_paper_validation.csv`
S2	Per-element accuracy breakdown (24 elements)	`output/paper_supplementary/S2_element_accuracy.csv`
S3	Consensus statistics and model contributions	`output/paper_supplementary/S3_consensus_stats.csv`
S4	Data completeness and capture rates	`output/paper_supplementary/S4_data_completeness.csv`