Text

AI-generated content may be incorrect.

**Article Title:**

**Automating Data Extraction from Scientific Literature and General PDF Files Using Large Language Models and KNIME: An Application in Toxicology**

**Authors:**

|  |
| --- |
| **José Teófilo Moreira-Filho\***  National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods, Division of Translational Toxicology, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA.  [teofilo.moreirafilho@nih.gov](mailto:teofilo.moreirafilho@nih.gov)  ORCID iD: <https://orcid.org/0000-0002-0777-280X> |
| **Dhruv Ranganath**  National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods, Division of Translational Toxicology, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA;  University of North Carolina, Chapel Hill, NC, USA.  [dhruv.ranganath@nih.gov](mailto:dhruv.ranganath@nih.gov) |
| **Ricardo S. Tieghi**  National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods, Division of Translational Toxicology, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA;  University of North Carolina, Chapel Hill, NC, USA.  ricardo.scheufentieghi@nih.gov |
| **Robert Patton**  Oak Ridge National Laboratory, Oak Ridge, TN, USA.  [pattonrm@ornl.gov](mailto:pattonrm@ornl.gov) |
| **Vicki Sutherland**  Johnson & Johnson Innovative Medicine, New Brunswick, NJ, USA.  [VSutherl@ITS.JNJ.com](mailto:VSutherl@ITS.JNJ.com) |
| **Charles Schmitt**  Division of Translational Toxicology, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA.  [charles.schmitt@nih.gov](mailto:charles.schmitt@nih.gov) |
| **Andrew A. Rooney**  Division of Translational Toxicology, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA.  [andrew.rooney@nih.gov](mailto:andrew.rooney@nih.gov) |
| **Jennifer Fostel**  Division of Translational Toxicology, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA.  [fostel@niehs.nih.gov](mailto:fostel@niehs.nih.gov) |
| **Vickie R. Walker**  Division of Translational Toxicology, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA.  [vickie.walker@nih.gov](mailto:vickie.walker@nih.gov) |
| **Trey Saddler**  Axle Informatics, Rockville, MD, USA.  [trey.saddler@nih.gov](mailto:trey.saddler@nih.gov) |
| **David Reif**  Predictive Toxicology Branch, Division of Translational Toxicology, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA.  [david.reif@nih.gov](mailto:david.reif@nih.gov) |
| **Kamel Mansouri\***  National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods, Division of Translational Toxicology, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA.  [kamel.mansouri@nih.gov](mailto:kamel.mansouri@nih.gov) |
| **Nicole Kleinstreuer**  National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods, Division of Translational Toxicology, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA.  [nicole.kleinstreuer@nih.gov](mailto:nicole.kleinstreuer@nih.gov) |

**Supporting Information:**

# Details on the Evaluation of Tabular Data Extraction Tools

To select a tool for extracting tabular data from PDF files, we tested six different tools: pdfplumber,1 tabula-py,2 Nougat,3 Marker,4 Docling,5 and Camelot.6 The evaluation criteria focused on accuracy (i.e., correct identification and extraction of tabular structures) and speed, as our goal was to integrate the selected tool with GROBID for efficient text and table extraction. For this evaluation, we used three peer-reviewed publications of toxicological studies.7–9 Below, we summarize the performance of each tool.

* **Pdfplumber** was unable to extract tables from the tested PDF files, eliminating it as a viable option for our implementation.
* **Tabula-py** was fast and extracted tables with good accuracy. It also provides an option to extract only tables, reducing post-processing efforts. These advantages made it a strong candidate for our final implementation.
* **Nougat** produces semi-structured output in a Markdown-like format. However, it lacks an option to extract only tables, requiring additional processing steps to isolate tabular content from the extracted text. Furthermore, Nougat's parsing speed is slow, making it a suboptimal choice for large-scale applications where efficiency is crucial.
* **Marker** (tested version 0.2.14) exhibited high accuracy in table extraction. However, similar to Nougat, it does not provide an option to extract only tables, necessitating post-processing to filter out the desired tabular data. Given our focus on efficiency, this additional step made it less favorable.
* **Docling** demonstrated high accuracy due to its reliance on TableFormer, a transformer-based model for table detection and extraction. It also provides an option to extract only tables. Although slower than Tabula-py, it is faster than Nougat, making it a reliable choice.
* **Camelot** is specifically designed for table extraction, but it demonstrated low accuracy, often misidentifying non-tabular elements as tables. Additionally, it failed to extract tables correctly in our tests. While fast, its lack of precision made it unsuitable for our use case.

Based on our evaluation, we selected Tabula-py and Docling for our final implementation. These tools provided a balance between speed and accuracy, meeting our objective of efficient and precise table extraction to complement GROBID’s text processing capabilities.

# Large Language Models

**Table S1**. LLMs context window, maximum output tokens and pricing.

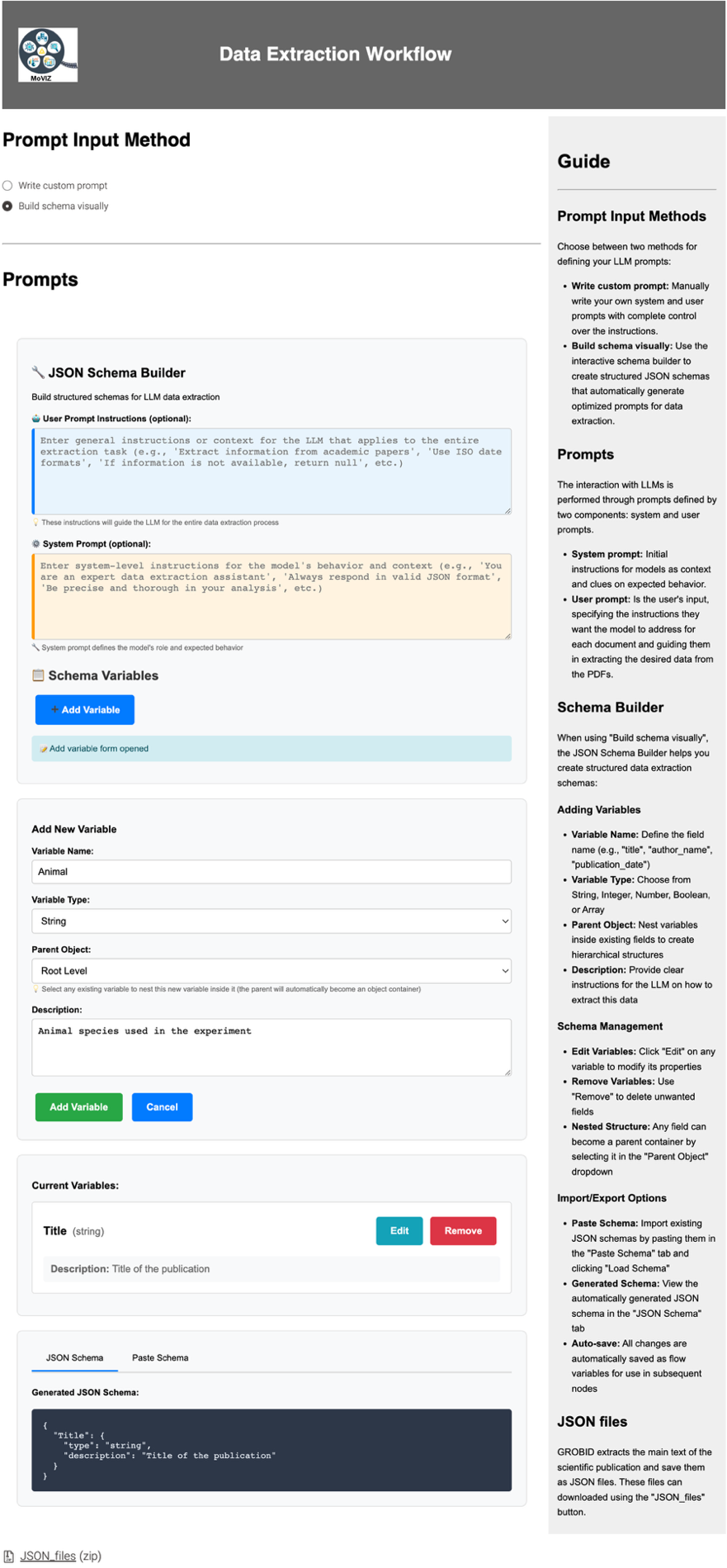
| **Model** | **Context token window** | **Max output tokens** | **Pricing of input** | **Pricing of output** | **Reference** |
| --- | --- | --- | --- | --- | --- |
| Claude 3 Opus - 20240229 | 200,000 | 4,096 | $15.00 / 1M tokens | $75.00 / 1M tokens | 10 |
| Claude 3.5 Sonnet - 20241022 | 200,000 | 8,192 | $3.00 / 1M tokens | $15.00 / 1M tokens | 10 |
| Claude 3 Haiku - 20240307 | 200,000 | 4,096 | $0.25 / 1M tokens | $1.25 / 1M tokens | 10 |
| GPT-4o | 128,000 | 16,284 | $2.50 / 1M tokens | $10.00 / 1M tokens | 11,12 |
| GPT-4o mini | 128,000 | 16,284 | $0.15 / 1M tokens | $0.60 / 1M output tokens | 11,12 |
| Llama 3.1 70B | 128,000 | 4,096 | - | - | 13 |
| Llama 3.1 8B | 128,000 | 4,096 | - | - | 13 |

# Supplementary Figures

A screenshot of a document

AI-generated content may be incorrect.

**Figure S1**. Image mode for scientific publications.



**Figure S2**. Screen of the user interface for prompt input using the “Build schema visually” option.

# Scientific Publications’ Extracted Variables

**Table S2**. Extracted variables in the text and image modes from scientific publications.

| **Variable** | **Description** |
| --- | --- |
| Title | Title of the scientific publication |
| Animal | Animal species used in the experiment |
| Age at treatment (Y/N) | Indicates if age at treatment initiation was reported |
| Age at treatment (Entity) | Specific age(s) of animals at initiation of treatment |
| Age at treatment (Source text) | Original text from the publication reporting animal age |
| Body weight at treatment? (Y/N) | Indicates if body weight at treatment initiation was reported |
| Body weight at treatment (Entity) | Specific body weight(s) of animals at treatment initiation |
| Body weigh at treatment (Source text) | Original text reporting body weight at treatment initiation |
| Number of animals per treatment group | Number of animals per group, indicating if reported as total, by sex, or evenly split |
| Route of administration | Method by which the substance was administered to animals |
| Dose | The administered dose of the test substance |
| Daily dosing? (Y/N) | Indicates if dosing occurred daily |
| Dose day | Days on which doses were administered |
| Animal checks during treatment? (Y/N) | Indicates if animals were monitored during treatment |
| Animal checks during treatment (Entity) | Description of animal health checks conducted |
| Animal checks during treatment? (Source text) | Original text describing animal checks |
| Body weights during treatment? (Y/N) | Indicates if body weights were measured during treatment |
| Body weights during treatment (Entity) | The extracted data on body weight measurements during treatment. |
| Body weights during treatment (Source text) | Original text reporting body weights during treatment |
| Food consumption during treatment? (Y/N) | Indicates if food consumption was measured during treatment |
| Food consumption during treatment (Entity) | Specific food consumption data reported |
| Food consumption during treatment (Source text) | Original text reporting food consumption |
| Sacrifice (Y/N) | Indicates if animals were sacrificed at end of experiment |
| Sacrifice (Entity) | Details of animal sacrifice procedure |
| Sacrifice (Source text) | Original text describing sacrifice procedure |
| Maternal body weight at sacrifice? (Y/N) | Indicates if maternal body weight at sacrifice was measured |
| Maternal body weight at sacrifice (Entity) | The extracted maternal body weight data at sacrifice |
| Maternal body weight at sacrifice (Source text) | Original text reporting maternal body weight at sacrifice |
| Fetal body weight at sacrifice, individual? (Y/N) | Indicates if individual fetal body weight was measured at sacrifice |
| Fetal body weight at sacrifice, individual (Entity) | The extracted individual fetal body weight data |
| Fetal body weight at sacrifice, individual (Source text) | Original text reporting individual fetal body weight |
| Fetal body weight at sacrifice, combined? (Y/N) | Indicates if combined fetal weight was measured |
| Fetal body weight at sacrifice, combined (Entity) | The extracted combined fetal body weight data |
| Fetal body weight at sacrifice, combined (Source text) | Original text reporting combined fetal weight |
| Uterine weight? (Y/N) | Indicates if uterine weight was measured |
| Uterine weight (Entity) | The extracted uterine weight data |
| Uterine weight (Source text) | Original text reporting uterine weight |
| Organ weights? (any in addition to uterus, Y/N) | Indicates if additional organ weights besides uterus were measured |
| Organ weights? (any in addition to uterus, entity) | The extracted organ weights data |
| Organ weights? (any in addition to uterus, source text) | Original text reporting organ weights |
| Pregnancy status (Y/N) | Indicates if pregnancy status was assessed |
| Pregnancy status (Entity) | The extracted pregnancy status information |
| Pregnancy status (Source text) | Original text reporting pregnancy status |
| Number/count of live fetuses? (Y/N) | Indicates if live fetus count was reported |
| Number/count of live fetuses (entity) | The extracted data of live fetuses count |
| Number/count of live fetuses (source text) | Original text reporting live fetus count |
| Number/count of dead fetuses? (Y/N) | Indicates if dead fetus count was reported |
| Number/count of dead fetuses (Entity) | The extracted data of dead fetuses count |
| Number/count of dead fetuses (Source text) | Original text reporting dead fetus count |
| Fetal sex (Y/N) | Indicates if fetal sex was reported |
| Fetal sex (Number of males and females, entity) | The extracted number of male and female fetuses |
| Fetal sex (Source text) | Original text reporting the number of male and female fetuses |
| Number/count of implantation sites? (Y/N) | Indicates if implantation site count was reported |
| Number/count of implantation sites (Entity) | The extracted number of implantation sites |
| Number/count of implantation sites (Source text) | Original text reporting implantation sites |
| Number/count of corpora lutea? (Y/N) | Indicates if corpora lutea count was reported |
| Number/count of corpora lutea (Entity) | The extracted number of corpora lutea |
| Number/count of corpora lutea (Source text) | Original text reporting corpora lutea |
| Number/count of resorptions? (Y/N) | Indicates if resorptions were reported |
| Number/count of resorptions (Entity) | The extracted number of resorptions |
| Number/count of resorptions (Source text) | Original text reporting resorptions |
| Placental evaluation? (Y/N) | Indicates if placental evaluations were conducted |
| Placental evaluation (Entity) | The extracted data on placental evaluation |
| Placental evaluation (Source text) | Original text reporting placental evaluations |
| Fetal exam, any type? (Y/N) | Indicates if any type of fetal examination was conducted |
| Fetal exam, any type (Entity) | The extracted data of fetal examination |
| Fetal exam, any type (Source text) | Original text describing fetal examinations |
| Fetal external examination? (Y/N) | Indicates if an external examination of fetuses was conducted |
| Fetal external examination (Entity) | The extracted data on fetal external examination |
| Fetal external examination (Source text) | Original text describing fetal external examination |
| Fetal visceral examination? (Y/N) | Indicates if a fetal visceral examination was conducted |
| Fetal visceral examination (Entity) | The extracted data of fetal visceral examination |
| Fetal visceral examination (Source text) | Original text related to fetal visceral examination |
| Fetal skeletal examination? (Y/N) | Indicates if a fetal skeletal examination was conducted |
| Fetal skeletal examination (Entity) | The extracted data of fetal skeletal examination |
| Fetal skeletal examination (Source text) | Original text related to fetal skeletal examination |

# Scientific Publications’ Prompts

**Table S3**. System and user prompts used in the text and image modes for data extraction from scientific publications.

| **Extraction mode** | **System prompt** | **User prompt** |
| --- | --- | --- |
| **Text** | You are a scientific assistant and toxicologist expert in reviewing scientific publications and extract and analyze specific relevant information out of PDFs in valid JSON format. Be careful with numbers. | Analyze the input PDFs and extract the specified variables to create a new JSON output. Follow these guidelines:  1. Output Format:  - Return only a valid JSON object.  - Do not include any additional text before or after the JSON output.  2. JSON Structure:  - Each variable becomes a key-value pair in the JSON object.  - Use the exact variable names provided in the list below.  3. Data Handling:  - Use 'N/A' for missing or null values.  - Ensure uniform formatting for similar types of variables.  - Include only the specified variables in the output.  - Combine related variables into a single value without nesting (e.g., combine "specie 1" and "specie 2" into a single "Animal" value).  4. Value Extraction Guidelines:  - For 'Y/N' variables: Use 'Y' if information is present, 'N' if absent.  - For entity variables: Provide specific values or ranges (e.g., '8 weeks', '200 grams').  - For source text variables: Provide the exact relevant text snippet from the input.  5. Specific Variable Handling:  - 'Animal': Provide species as a single string (e.g., 'Rat', 'Mouse'). Separate multiple species with commas.  - 'Number of animals per treatment group': Specify as '10M and 10F' or '20 overall evenly split' if applicable.  - 'Dose': Include amount and units (e.g., '10 mg/kg').  - 'Dose day': Specify the exact days of administration.  6. Complex Data Handling:  - For variables with multiple related pieces of information, combine them into a coherent, comma-separated string.  - If information is spread across multiple sections in the input, consolidate it into a single, comprehensive value.  7. Consistency and Accuracy:  - Maintain consistent formatting across similar variables.  - Double-check extracted values against the source text to ensure accuracy.  8. Error Handling and Reporting:  - If unable to extract a value due to complexity or lack of clear information, use 'Unable to determine' instead of 'N/A'.  Examples of Correct Extractions:  - "Animal": "Sprague-Dawley rats"  - "Age at treatment (Entity)": "8-10 weeks"  - "Body weight at treatment (Entity)": "180-220 grams"  - "Dose": "10 mg/kg, 50 mg/kg, 100 mg/kg"  - "Dose day": "GD 6-15"  - "Number of animals per treatment group": "25 females per dose group"  - "Body weight at treatment? (entity)”: “Rats: 160 g, Mice: 20 g”  Variables to extract:  [  'Title',  'Animal',  'Age at treatment (Y/N)',  'Age at treatment (Entity)',  'Age at treatment (Source text)',  'Body weight at treatment? (Y/N)',  'Body weight at treatment (Entity)',  'Body weigh at treatment (Source text)',  'Number of animals per treatment group (note if this is given as #M and #F, or # overall evenly split between the two)',  'Route of administration',  'Dose',  'Daily dosing? (Y/N)',  'Dose day',  'Animal checks during treatment? (Y/N)',  'Animal checks during treatment (Entity)',  'Animal checks during treatment? (Source text)',  'Body weights during treatment? (Y/N)',  'Body weights during treatment (Entity)',  'Body weights during treatment (Source text)',  'Food consumption during treatment? (Y/N)',  'Food consumption during treatment (Entity)',  'Food consumption during treatment (Source text)',  'Sacrifice (Y/N)',  'Sacrifice (Entity)',  'Sacrifice (Source text)',  'Maternal body weight at sacrifice? (Y/N)',  'Maternal body weight at sacrifice (Entity)',  'Maternal body weight at sacrifice (Source text)',  'Fetal body weight at sacrifice, individual? (Y/N)',  'Fetal body weight at sacrifice, individual (Entity)',  'Fetal body weight at sacrifice, individual (Source text)',  'Fetal body weight at sacrifice, combined? (Y/N)',  'Fetal body weight at sacrifice, combined (Entity)',  'Fetal body weight at sacrifice, combined (Source text)',  'Uterine weight? (Y/N)',  'Uterine weight (Entity)',  'Uterine weight (Source text)',  'Organ weights? (any in addition to uterus, Y/N)',  'Organ weights? (any in addition to uterus, entity)',  'Organ weights? (any in addition to uterus, source text)',  'Pregnancy status (Y/N)',  'Pregnancy status (Entity)',  'Pregnancy status (Source text)',  'Number/count of live fetuses? (Y/N)',  'Number/count of live fetuses (entity)',  'Number/count of live fetuses (source text)',  'Number/count of dead fetuses? (Y/N)',  'Number/count of dead fetuses (Entity)',  'Number/count of dead fetuses (Source text)',  'Fetal sex (Y/N)',  'Fetal sex (Number of males and females, entity)',  'Fetal sex (Source text)',  'Number/count of implantation sites? (Y/N)',  'Number/count of implantation sites (Entity)',  'Number/count of implantation sites (Source text)',  'Number/count of corpora lutea? (Y/N)',  'Number/count of corpora lutea (Entity)',  'Number/count of corpora lutea (Source text)',  'Number/count of resorptions? (Y/N)',  'Number/count of resorptions (Entity)',  'Number/count of resorptions (Source text)',  'Placental evaluation? (Y/N)',  'Placental evaluation (Entity)',  'Placental evaluation (Source text)',  'Fetal exam, any type? (Y/N)',  'Fetal exam, any type (Entity)',  'Fetal exam, any type (Source text)',  'Fetal external examination? (Y/N)',  'Fetal external examination (Entity)',  'Fetal external examination (Source text)',  'Fetal visceral examination? (Y/N)',  'Fetal visceral examination (Entity)',  'Fetal visceral examination (Source text)',  'Fetal skeletal examination? (Y/N)',  'Fetal skeletal examination (Entity)',  'Fetal skeletal examination (Source text)'] |
| **Image** | You are a scientific assistant and toxicologist expert in reviewing scientific publications and extract and analyze specific relevant information out of PDFs in valid JSON format. Be careful with numbers. | Analyze the input PDFs and extract the specified variables to create a new JSON output. Follow these guidelines:  1. Output Format:  - Return only a valid JSON object.  - Do not include any additional text before or after the JSON output.  2. JSON Structure:  - Each variable becomes a key-value pair in the JSON object.  - Use the exact variable names provided in the list below.  3. Data Handling:  - Use 'N/A' for missing or null values.  - Ensure uniform formatting for similar types of variables.  - Include only the specified variables in the output.  - Combine related variables into a single value without nesting (e.g., combine "specie 1" and "specie 2" into a single "Animal" value).  4. Value Extraction Guidelines:  - For 'Y/N' variables: Use 'Y' if information is present, 'N' if absent.  - For entity variables: Provide specific values or ranges (e.g., '8 weeks', '200 grams').  - For source text variables: Provide the exact relevant text snippet from the input.  5. Specific Variable Handling:  - 'Animal': Provide species as a single string (e.g., 'Rat', 'Mouse'). Separate multiple species with commas.  - 'Number of animals per treatment group': Specify as '10M and 10F' or '20 overall evenly split' if applicable.  - 'Dose': Include amount and units (e.g., '10 mg/kg').  - 'Dose day': Specify the exact days of administration.  6. Complex Data Handling:  - For variables with multiple related pieces of information, combine them into a coherent, comma-separated string.  - If information is spread across multiple sections in the input, consolidate it into a single, comprehensive value.  7. Consistency and Accuracy:  - Maintain consistent formatting across similar variables.  - Double-check extracted values against the source text to ensure accuracy.  8. Error Handling and Reporting:  - If unable to extract a value due to complexity or lack of clear information, use 'Unable to determine' instead of 'N/A'.  Examples of Correct Extractions:  - "Animal": "Sprague-Dawley rats"  - "Age at treatment (Entity)": "8-10 weeks"  - "Body weight at treatment (Entity)": "180-220 grams"  - "Dose": "10 mg/kg, 50 mg/kg, 100 mg/kg"  - "Dose day": "GD 6-15"  - "Number of animals per treatment group": "25 females per dose group"  - "Body weight at treatment? (entity)”: “Rats: 160 g, Mice: 20 g”  Variables to extract:  [  'Title',  'Animal',  'Age at treatment (Y/N)',  'Age at treatment (Entity)',  'Age at treatment (Source text)',  'Body weight at treatment? (Y/N)',  'Body weight at treatment (Entity)',  'Body weigh at treatment (Source text)',  'Number of animals per treatment group (note if this is given as #M and #F, or # overall evenly split between the two)',  'Route of administration',  'Dose',  'Daily dosing? (Y/N)',  'Dose day',  'Animal checks during treatment? (Y/N)',  'Animal checks during treatment (Entity)',  'Animal checks during treatment? (Source text)',  'Body weights during treatment? (Y/N)',  'Body weights during treatment (Entity)',  'Body weights during treatment (Source text)',  'Food consumption during treatment? (Y/N)',  'Food consumption during treatment (Entity)',  'Food consumption during treatment (Source text)',  'Sacrifice (Y/N)',  'Sacrifice (Entity)',  'Sacrifice (Source text)',  'Maternal body weight at sacrifice? (Y/N)',  'Maternal body weight at sacrifice (Entity)',  'Maternal body weight at sacrifice (Source text)',  'Fetal body weight at sacrifice, individual? (Y/N)',  'Fetal body weight at sacrifice, individual (Entity)',  'Fetal body weight at sacrifice, individual (Source text)',  'Fetal body weight at sacrifice, combined? (Y/N)',  'Fetal body weight at sacrifice, combined (Entity)',  'Fetal body weight at sacrifice, combined (Source text)',  'Uterine weight? (Y/N)',  'Uterine weight (Entity)',  'Uterine weight (Source text)',  'Organ weights? (any in addition to uterus, Y/N)',  'Organ weights? (any in addition to uterus, entity)',  'Organ weights? (any in addition to uterus, source text)',  'Pregnancy status (Y/N)',  'Pregnancy status (Entity)',  'Pregnancy status (Source text)',  'Number/count of live fetuses? (Y/N)',  'Number/count of live fetuses (entity)',  'Number/count of live fetuses (source text)',  'Number/count of dead fetuses? (Y/N)',  'Number/count of dead fetuses (Entity)',  'Number/count of dead fetuses (Source text)',  'Fetal sex (Y/N)',  'Fetal sex (Number of males and females, entity)',  'Fetal sex (Source text)',  'Number/count of implantation sites? (Y/N)',  'Number/count of implantation sites (Entity)',  'Number/count of implantation sites (Source text)',  'Number/count of corpora lutea? (Y/N)',  'Number/count of corpora lutea (Entity)',  'Number/count of corpora lutea (Source text)',  'Number/count of resorptions? (Y/N)',  'Number/count of resorptions (Entity)',  'Number/count of resorptions (Source text)',  'Placental evaluation? (Y/N)',  'Placental evaluation (Entity)',  'Placental evaluation (Source text)',  'Fetal exam, any type? (Y/N)',  'Fetal exam, any type (Entity)',  'Fetal exam, any type (Source text)',  'Fetal external examination? (Y/N)',  'Fetal external examination (Entity)',  'Fetal external examination (Source text)',  'Fetal visceral examination? (Y/N)',  'Fetal visceral examination (Entity)',  'Fetal visceral examination (Source text)',  'Fetal skeletal examination? (Y/N)',  'Fetal skeletal examination (Entity)',  'Fetal skeletal examination (Source text)'] |

# General PDFs’ (Safety Data Sheets) Extracted Variables

**Table S4**. Extracted variables in the text and image modes from safety data sheets (general PDFs).

| **Variable** | **Description** |
| --- | --- |
| Formulation name | Name of the product formulation |
| CAS Numbers | Chemical Abstracts Service (CAS) numbers of the components in the formulation |
| Contents (W/W) | Percentage composition (weight/weight) of each component in the formulation |
| Oral Type of Value | Type of toxicity value for oral exposure (e.g., LD50) |
| Oral Species | Species used in oral toxicity testing |
| Oral Value | Measured toxicity value for oral exposure (e.g., LD50 in mg/kg) |
| Oral Method | Test guideline used for oral toxicity assessment (e.g., OECD 401) |
| Inhalation Type of Value | Type of toxicity value for inhalation exposure (e.g., LC50) |
| Inhalation Species | Species used in inhalation toxicity testing |
| Inhalation Value | Measured toxicity value for inhalation exposure (e.g., LC50 in mg/L) |
| Inhalation Method | Test guideline used for inhalation toxicity assessment (e.g., OECD 403) |
| Dermal Type of Value | Type of toxicity value for dermal exposure (e.g., LD50) |
| Dermal Species | Species used in dermal toxicity testing |
| Dermal Value | Measured toxicity value for dermal exposure (e.g., LD50 in mg/kg) |
| Dermal Method | Test guideline used for dermal toxicity assessment (e.g., OECD 402) |
| Assessment of Irritating Effects | Summary of irritation effects on skin and eyes |
| Skin Species | Species used in skin irritation testing |
| Skin Result | Observed skin irritation result (e.g., non-irritant) |
| Skin Method | Test guideline used for skin irritation assessment (e.g., OECD 404) |
| Eye Species | Species used in eye irritation testing |
| Eye Result | Observed eye irritation result (e.g., moderately irritating) |
| Eye Method | Test guideline used for eye irritation assessment (e.g., OECD 405) |
| Assessment of Sensitization | Summary of sensitization effects from skin exposure |
| Guinea Pig Maximization Test Species | Species used in the guinea pig maximization test |
| Guinea Pig Maximization Test Result | Observed result of guinea pig maximization test for skin sensitization |
| Mouse Local Lymph Node Assay (LLNA) Result | Observed result of LLNA test for sensitization |
| Mouse Local Lymph Node Assay (LLNA) Method | Test guideline used for LLNA sensitization assessment (e.g., OECD 429) |

# General PDFs’ (Safety Data Sheets) Prompts

**Table S5**. System and user prompts used in the text and image modes for data extraction from safety data sheets (general PDFs).

| **Extraction mode** | **System prompt** | **User prompt** |
| --- | --- | --- |
| **Text** | You are a scientific assistant and toxicologist expert in reviewing reports and extract and analyze specific relevant information out of PDFs in valid JSON format. Be careful with numbers. | Analyze the input markdowns and extract the specified variables to create a new JSON output. Follow these guidelines:  Extract data related to a Formulation and its Acute Toxicity/Effects from input markdown files and output as a structured flat JSON format. The information includes values for specific formulation composition, toxicological tests, and methodologies. It does not require nested JSON structures.  ## Steps:  1. Read the file, then identify and extract the relevant sections from the markdown text below.  2. Parse the necessary values as described for each section.  3. Structure the extracted data as flat key-value pairs in a JSON format.  4. For chemical names and CAS numbers, if more than one, combine into a single value format where they are separated by a comma (,).  5. Iterate through the remaining files, repeating steps 1 through 4  ## Sections to Extract:  ### 1. Identification  • Formulation name  ### 3. Composition / Information on Ingredients  • CAS Number  • Content (W/W)  ### 11. Toxicological Information  • Acute Toxicity/Effects  • Oral: Type of Value, Species, Value, and Method.  • Inhalation: Type of Value, Species, Value, and Method.  • Dermal: Type of Value, Species, Value, and Method.  • Irritation / Corrosion  • Assessment of Irritating Effects: Overall assessment.  • Specific Tests  • Skin: Species, Result, and Method.  • Eye: Species, Result, and Method.  • Sensitization: Assessment of sensitization.  • Guinea Pig Maximization Test: Species and Result.  • Mouse Local Lymph Node Assay: Result and Method  # Output Format  The output should be a flat JSON object with key-value pairs where each toxicological measure's name is concatenated with its detail type. Example keys: "Oral Type of Value", "Oral Species", "Oral Value", etc.  Example Output:  {  "Formulation name": "Acrobat CT",  "CAS Numbers": "23564-05-8, 175013-18-0, 120068-37-3, 56-81-5, 9043-30-5, 1328-53-6",  "Contents (W/W)":"15.0 - 20.0%, 1.0 - 3.0%, 20.0 - 25.0%, 3.0 - 5.0%, 0.1 - 0.3%, 1.0 - 3.0"  "Oral Type of Value": "LD50",  "Oral Species": "rat (male)",  "Oral Value": "> 4,000 mg/kg",  "Oral Method": "OECD Guideline 401",  "Inhalation Type of Value": "LC50",  "Inhalation Species": "rat (male/female)",  "Inhalation Value": "> 5.44 mg/l",  "Inhalation Method": "OECD Guideline 403",  "Dermal Type of Value": "LD50",  "Dermal Species": "rat (male/female)",  "Dermal Value": "> 20,000 mg/kg",  "Dermal Method": "OECD Guideline 402",  "Assessment of Irritating Effects": "Not irritating to the skin. Eye contact causes irritation.",  "Skin Species": "rabbit",  "Skin Result": "non-irritant",  "Skin Method": "OECD Guideline 404",  "Eye Species": "rabbit",  "Eye Result": "moderately irritating",  "Eye Method": "OECD Guideline 405",  "Assessment of Sensitization": "Sensitization after skin contact possible.",  "Guinea Pig Maximization Test Species": "guinea pig",  "Guinea Pig Maximization Test Result": "Skin sensitizing effects were not observed in animal studies."  "Mouse Local Lymph Node Assay (LLNA) Result": "Non-sensitizing",  "Mouse Local Lymph Node Assay (LLNA) Method": "OECD Guideline 429"  }  # Notes  • Ensure data integrity by correctly matching the extracted data with the appropriate labels.  • Use placeholders where applicable for sections if they are not available.  • Maintain consistency in naming and structure for ease of processing.  # Important  • Ignore the example input and output provided in this instruction.  • Ignore sections in the markdown which are not relevant to the variables, focusing only on the important sections, which are: 1 (Identification), 3 (Composition / Information on ingredients) and 11 (Toxicological Information)  • If a section or value is not available in the the following markdown file, use placeholders "N/A" in the the key-value pair.  • Extract information only from the following markdown files.  • Once you complete the extraction for a markdown, ignore your answers and start from scratch. |
| **Image** | You are a scientific assistant and toxicologist expert in reviewing reports and extracting and analyzing specific relevant information from images. Be careful with numbers. | Analyze the input images and extract the specified variables to create a new JSON output. Follow these guidelines:  Extract data related to a Formulation and its Acute Toxicity/Effects from input images and output it in a structured flat JSON format. The information includes values for specific formulation composition, toxicological tests, and methodologies. It does not require nested JSON structures.  ## Sections to Extract:   1. \*\*Identification\*\*    * Formulation name 2. \*\*Composition / Information on Ingredients\*\*    * CAS Number    * Content (W/W) 3. \*\*Acute Toxicity/Effects\*\*    * Oral: Type of Value, Value Unit, Species, Value, and Method.    * Inhalation: Type of Value, Value Unit, Species, Value, and Method.    * Dermal: Type of Value, Value Unit, Species, Value, and Method. 4. \*\*Irritation/Corrosion\*\*    * \*\*Assessment of Irritating Effects:\*\* Overall assessment. 5. \*\*Specific Tests \*\*    * \*\*Skin:\*\* Species, Result, and Method.    * \*\*Eye:\*\* Species, Result, and Method.    * \*\*Sensitization:\*\* Assessment of sensitization.    * \*\*Guinea Pig Maximization Test:\*\* Species and Result.    * \*\*Mouse Local Lymph Node Assay:\*\* Result and Method.   ### Steps:   * Identify and extract the relevant sections from the images. * Parse the necessary values as described for each section. * Structure the extracted data as flat key-value pairs in a JSON format. * For chemical names and CAS numbers, if more than one, split into separate columns (CASRN\_1, CASRN\_2….   # Output Format  The output should be a flat JSON object with key-value pairs where each toxicological measure's name is concatenated with its detail type. Example keys: "Oral Type of Value", "Oral Species", "Oral Value", etc.  # Examples  Example Input:  ## Acrobat CT  ### 3. Composition / Information on Ingredients  thiophanate-methyl  CAS Number: 23564-05-8  Content (W/W): 15.0 - 20.0%  Synonym: Carbamic acid, [1,2-phenylenebis(iminocarbamothioyl)]bis-, dimethyl  ester  Pyraclostrobin  CAS Number: 175013-18-0  Content (W/W): 1.0 - 3.0%  Synonym: No data available.  Fipronil  CAS Number: 120068-37-3  Content (W/W): 20.0 - 25.0%  Synonym: Fipronil  glycerol  CAS Number: 56-81-5  Content (W/W): 3.0 - 5.0%  Synonym: 1,2,3-Propanetriol; Glycerol  (POLYMER) Poly(oxy-1,2-ethanediyl), .alpha.-isotridecyl-.omega.-hydroxy  CAS Number: 9043-30-5  Content (W/W): 0.1 - 0.3%  Synonym: .alpha.-Isotridecyl-.omega.-hydroxypoly(oxy-1,2-ethanediyl)  C.I. Pigment Green 7  CAS Number: 1328-53-6  Content (W/W): 1.0 - 3.0%  Synonym: C.I. 74260  #### Acute Toxicity/Effects  Oral  Type of value: LD50  Species: rat (male)  Value: > 4,000 mg/kg (OECD Guideline 401)  Inhalation  Type of value: LC50  Species: rat (male/female)  Value: > 5.44 mg/l (OECD Guideline 403)  Dermal  Type of value: LD50  Species: rat (male/female)  Value: > 20,000 mg/kg (OECD Guideline 402)  Irritation / corrosion  Assessment of irritating effects: Not irritating to the skin. Eye contact causes irritation.  Skin  Species: rabbit  Result: non-irritant  Method: OECD Guideline 404  Eye  Species: rabbit  Result: moderately irritating  Method: OECD Guideline 405  Sensitization  Assessment of sensitization: Sensitization after skin contact possible.  Guinea pig maximization test  Species: guinea pig  Result: Skin sensitizing effects were not observed in animal studies.  Example Output:  {  "Formulation name": "Acrobat CT",  "CAS Number 1": "23564-05-8”,  “CAS Number 2”: “175013-18-0”,  “CAS Number 3”: “120068-37-3”,  “CAS Number 4”: “56-81-5”,  “CAS Number 5”: “9043-30-5”,  “CAS Number 6”: “1328-53-6",  “CAS Number 7”: “N/A",  “CAS Number 8”: “N/A",  “CAS Number 9”: “N/A",  "Contents (W/W\_1":"15.0 - 20.0%,  “Contents (W/W\_2”: “1.0 - 3.0%”,  “Contents (W/W\_3”: “20.0 - 25.0%”,  “Contents (W/W\_4”: “3.0 - 5.0%”,  “Contents (W/W\_5”: “0.1 - 0.3%”,  “Contents (W/W\_6”: “1.0 - 3.0%",  “Contents (W/W\_7”: “N/A”  “Contents (W/W\_8”: “N/A”  “Contents (W/W\_9”: “N/A”  "Oral Type of Value": "LD50",  "Oral Species": "rat (male)",  "Oral Value": "> 4,000",  "Oral Value Unit": "mg/kg",  "Oral Method": "OECD Guideline 401",  "Inhalation Type of Value": "LC50",  "Inhalation Species": "rat (male/female)",  "Inhalation Value": "> 5.44",  "Inhalation Value Unit": "mg/l",  "Inhalation Method": "OECD Guideline 403",  "Dermal Type of Value": "LD50",  "Dermal Species": "rat (male/female)",  "Dermal Value": "> 20,000",  "Dermal Value Unit": "mg/kg",  "Dermal Method": "OECD Guideline 402",  "Assessment of Irritating Effects": "Not irritating to the skin. Eye contact causes irritation.",  "Skin Species": "rabbit",  "Skin Result": "non-irritant",  "Skin Method": "OECD Guideline 404",  "Eye Species": "rabbit",  "Eye Result": "moderately irritating",  "Eye Method": "OECD Guideline 405",  "Assessment of Sensitization": "Sensitization after skin contact possible.",  "Guinea Pig Maximization Test Species": "guinea pig",  "Guinea Pig Maximization Test Result": "Skin sensitizing effects were not observed in animal studies."  "Mouse Local Lymph Node Assay (LLNA) Result": "Non-sensitizing",  "Mouse Local Lymph Node Assay (LLNA) Method": "OECD Guideline 429"  }  # Notes   * Ensure data integrity by correctly matching the extracted data with the appropriate labels. * Use placeholders where applicable for sections if they are not available. * Maintain consistency in naming and structure for ease of processing.   # Important   * Ignore the example input and output provided in this instruction. * If a section or value is not available in the following images, use placeholders "N/A" in the key-value pair. * Extract information only from the following images. |

**References**

1 pdfplumber. <https://github.com/jsvine/pdfplumber> (accessed 5 Dec2024).

2 tabula-py. <https://github.com/chezou/tabula-py/tree/master> (accessed 5 Dec2024).

3 Blecher L, Cucurull G, Scialom T, Stojnic R. Nougat: Neural Optical Understanding for Academic Documents. *arXiv* 2023. doi:10.48550/arxiv.2308.13418.

4 marker. <https://github.com/VikParuchuri/marker> (accessed 5 Dec2024).

5 Auer C, Lysak M, Nassar A, Dolfi M, Livathinos N, Vagenas P *et al.* Docling Technical Report. *arXiv* 2024. doi:10.48550/arxiv.2408.09869.

6 Camelot. <https://github.com/atlanhq/camelot> (accessed 5 Dec2024).

7 Xuying W, Jiangbo Z, Yuping Z, Xili M, Yiwen Z, Tianbao Z *et al.* Effect of Astragaloside IV on the General and Peripartum Reproductive Toxicity in Sprague-Dawley Rats. *Int J Toxicol* 2010; **29**: 505–516.

8 Heuschmid FF, Schneider S, Schuster P, Lauer B, Ravenzwaay B van. Developmental toxicity of polyethylene glycol-g-polyvinyl alcohol grafted copolymer in rats and rabbits. *Food Chem Toxicol* 2013; **51**: S14–S23.

9 Althali NJ, Hassan AM, Abdel-Wahhab MA. Effect of grape seed extract on maternal toxicity and in utero development in mice treated with zearalenone. *Environ Sci Pollut Res* 2019; **26**: 5990–5999.

10 Anthropic Models. <https://docs.anthropic.com/en/docs/about-claude/models> (accessed 19 Dec2024).

11 OpenAI Models. <https://platform.openai.com/docs/models/gp#gpt-4o> (accessed 19 Dec2024).

12 OpenAI Pricing. <https://openai.com/api/pricing/> (accessed 19 Dec2024).

13 Llama 3.1. <https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_1> (accessed 19 Dec2024).