

Single Day Event

Overcoming Challenges in Automating Report Creation for Clinical Trials Using NLP & DL

*Surabhi Sundar,
Karthik Palani
Pfizer*

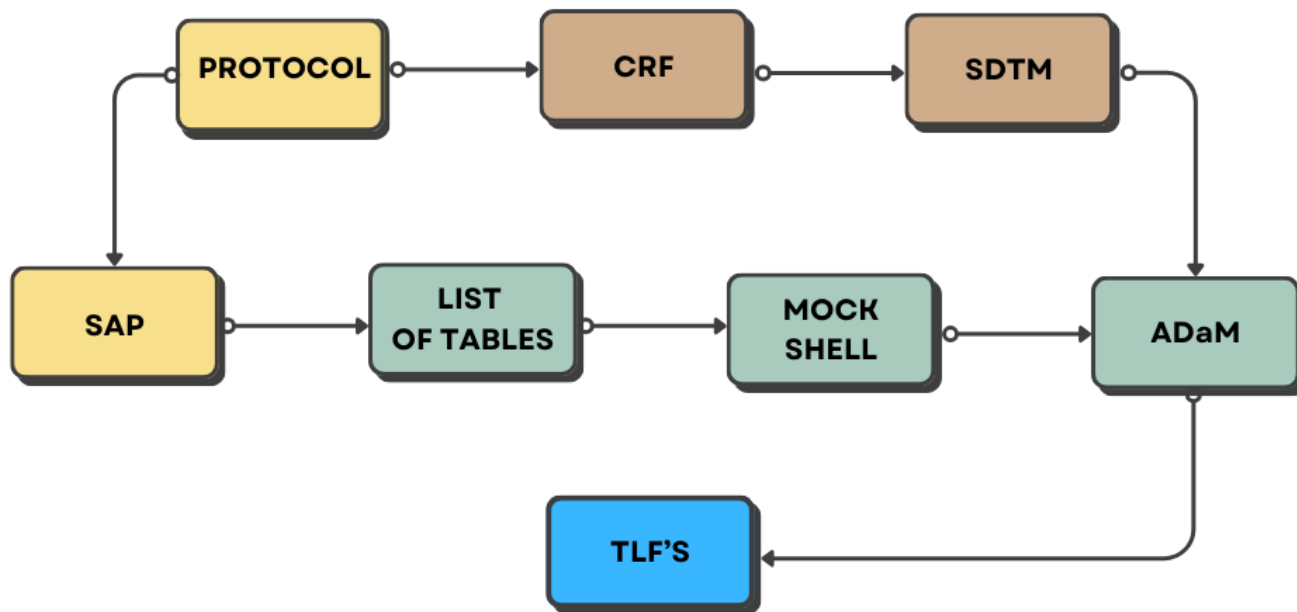


Disclaimer

We confirm that, the opinions and thoughts discussed in this presentation are subject to our own independent views and are not subject to the opinions of the organization that we represent.



Report Generation Process: A Glimpse





Report Generation Process: A Glimpse

Enhance Effectiveness

01

Streamline the Process

02

Reduce Dependency

03

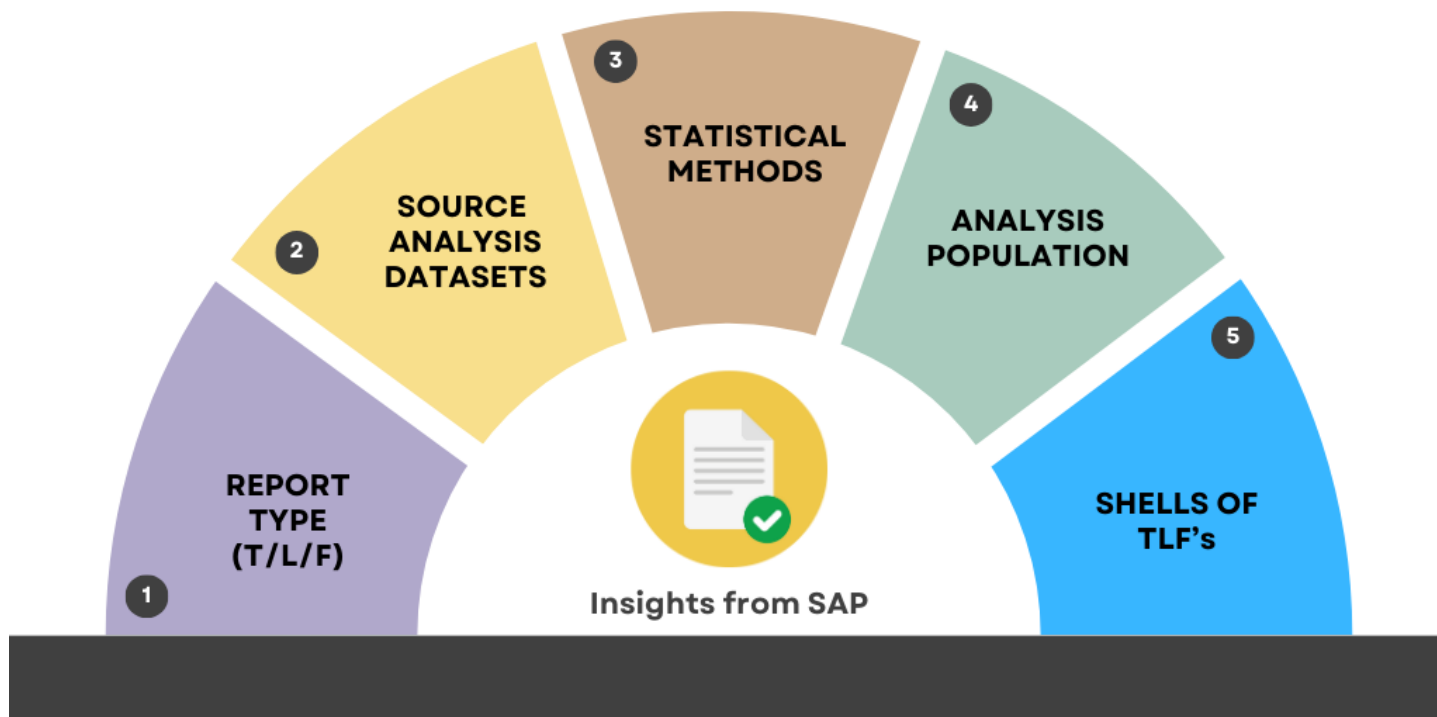
Time and Resources

04



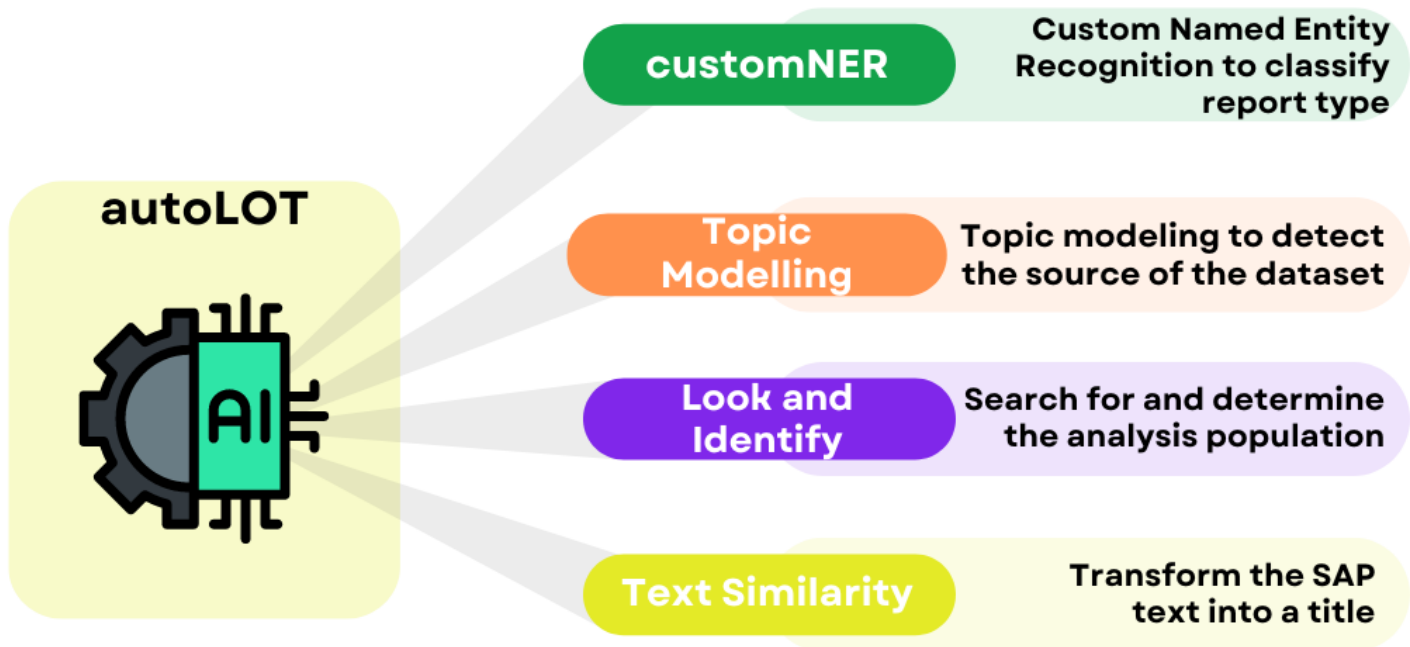


Revamping Report Creation Procedure





autoLOT Components



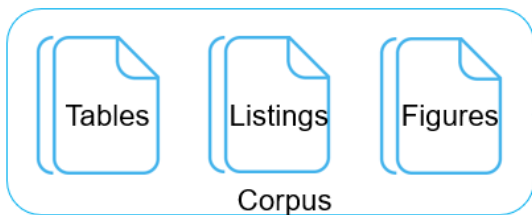


autoLOT Components

customNER

**Custom Named Entity
Recognition to classify
report type**

NER is an information extraction technique to identify and classify named entities in text



The number and percentage of subjects in the
Safety Analysis Set
(SAP text)



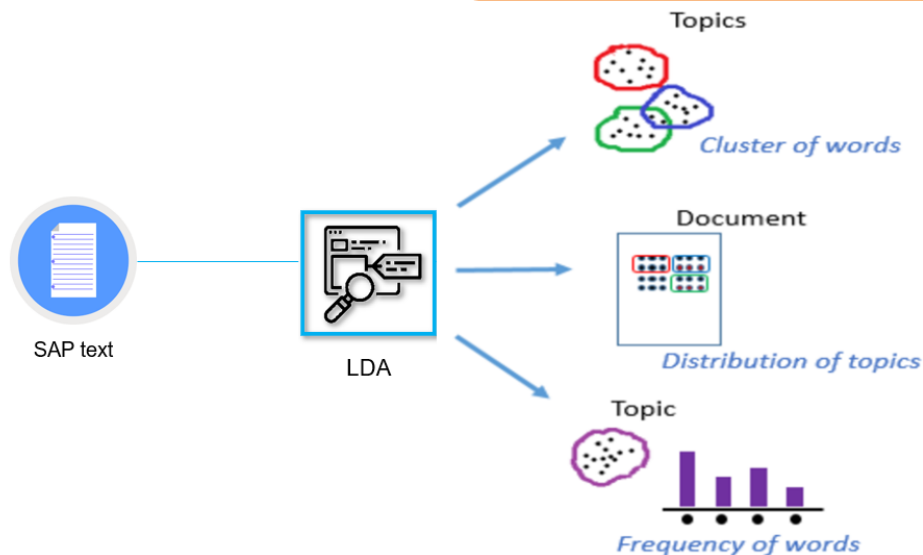
Custom
NER

[Tables]
The **number and percentage** of
subjects in the Safety Analysis Set

Topic Modelling

Topic modeling to detect the source of the dataset

Topic Modelling is a text mining technique that analyzes the semantic patterns in text to make sense of unstructured data without the need for predefined training data.



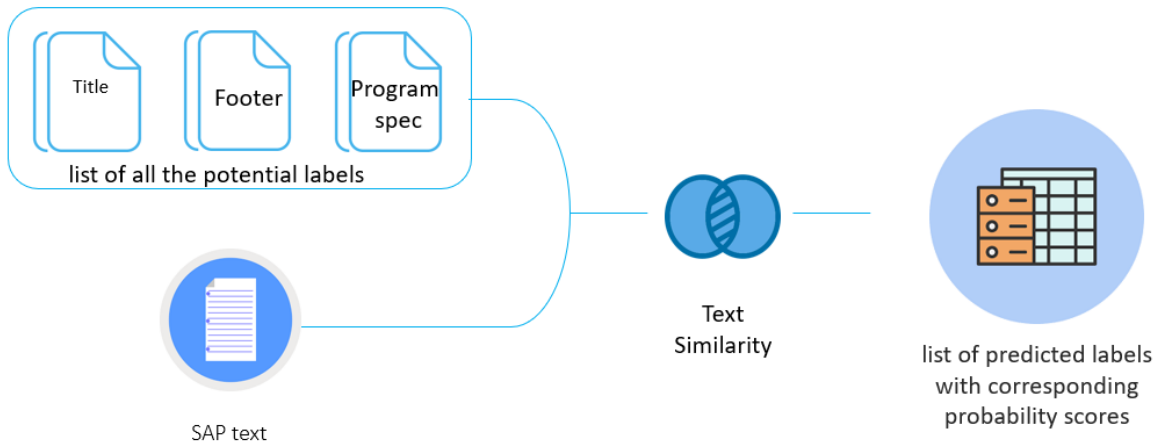


autoLOT Components

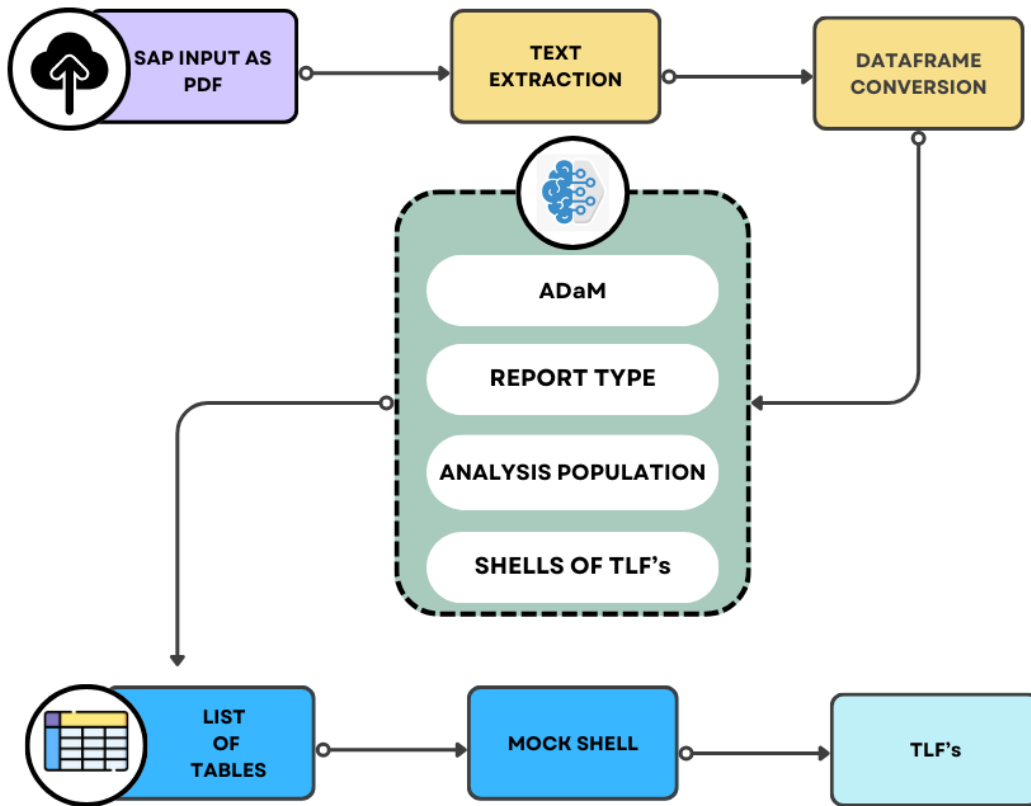
Text Similarity

Transform the SAP text into a title

Text Similarity is the process of comparing a piece of text with another and finding the similarity between them.



Discover the Functionality





autoLOT: An Illustrative Showcase

Extracting text from SAP

autoLOT

TLF's

autoLOT

Upload a SAP file for analysis.

Drag and drop file here
Limit 200MB per file • PDF

Browse files

SAP_001.pdf 3.0MB

×

☒ Structured SAP text

SAP text

▼



autoLOT: An Illustrative Showcase

Transform unstructured text data into a structured data frame

The screenshot displays the autoLOT web interface. On the left, a sidebar shows the Pfizer logo and a file list with 'autoLOT' and 'TLF's'. The main area features a file upload section with a 'Drag and drop file here' instruction, a 'Limit 200MB per file • PDF' note, and a 'Browse files' button. Below this, a file named 'SAP_001.pdf' (3.0MB) is shown. A checkbox labeled 'Structured SAP text' is checked. The structured output is displayed in a table with two columns: 'Title' and 'Text'. The table contains a list of sections from a Clinical Study Protocol (CSP), including '6. ANALYSES AND SUMMARIES' and its sub-sections. The text in the 'Text' column is partially obscured by redacted areas.

Title	Text
6. ANALYSES AND SUMMARIES	
6.1. Primary Endpoint(s)	the following pk parameters (including both primary and tertiary endpoints) will be summarized by hepatic function
6.2. Secondary Endpoint(s)	a set of summary tables split by hepatic function group will be produced to evaluate any potential risk associated with
6.2.1. Treatment and Disposition of Subjects	participant evaluation groups will show subject disposition, frequency counts will be supplied for participants discontinued
6.2.2. Demographic and Clinical Examination Data	a summary of demographic data will be provided for age, race, weight, body mass index, and height, each will be summarized
6.2.3. Discontinuation(s)	participant discontinuations will be detailed and summarized by hepatic function group, data will be reported in accordance with
6.2.4. Adverse Events	adverse events will be reported in accordance with the sponsor reporting standards by hepatic function group.
6.2.5. Laboratory Data	laboratory data will be listed and summarized in accordance with the sponsor reporting standards, the baseline measurement
6.2.6. Vital Signs Data	the baseline measurement (for blood pressure and pulse rate) is the last pre-dose measurement on day 1, absolute and relative
6.2.7. ECG Data	the baseline measurement is the last pre-dose measurement on day 1, absolute and changes from baseline for the entire study
6.2.8. Concomitant Treatments	all concomitant medication(s) as well as non-drug treatment(s) will be provided in the listings.

1 to 11 of 14 < > Page 1 of 2 > >



autoLOT: An Illustrative Showcase

Widgets to choose from

The screenshot displays the autoLOT interface with the Pfizer logo at the top left. Below the logo, there are tabs for 'autoLOT' and 'TLF's'. The main section is titled 'Reports of' and contains a 'Domains' checkbox labeled 'Selected TLF'. Two callout boxes are shown, each connected to a dropdown menu in the interface by a blue line.

Document Type

- List of Tables
- List of Tables
- Mock shells

Therapeutic areas

- Phase 1 and Research
- Phase 1 and Research
- Full Development
- Oncology



autoLOT: An Illustrative Showcase

Breaking apart into specific domains

autoLOT

TLF's

Document Type

List of Tables

Therapeutic areas

Phase 1 and Research

Go to

6.1. Primary End... x

6.2.2. Demograp... x

6.2.6. Vital Signs ... x

6. ANALYSES AND SUMMARIES

6.2. Secondary Endpoint(s)

6.2.1. Treatment and Disposition of S...

6.2.3. Discontinuation(s)

Reports of

Domains

6.1. Primary Endpoint(s)

6.2.2. Demographic and Clinical Examination Data

6.2.6. Vital Signs Data

☐ Selected TLF

Analysis of specific domains

Domains

6.1. Primary Endpoint(s)

SAP text

the following **pk parameters** including both primary and tertiary endpoints) will be summarized by hepatic function group **box and whisker plots** or individual subject **parameters** (aucinf, aucinf,u, cci cmax and cmax,u) be presented by hepatic function group and overlaid with geometric means. presentations for pf-06835919 concentrations will include: - a listing of all concentrations sorted by hepatic function group (present in heading), subject id and nominal time post-dose, the concentration listing will also include the actual times, deviations from the nominal time will be given in a separate listing. - a summary of concentrations by hepatic function group and nominal time post-dose, where the set of statistics will include n, mean, median, standard deviation, coefficient of variation (cv), minimum, maximum and the number of concentrations above the lower limit of quantification. - median concentrations time plots (on both linear and semi-log scales) against nominal time post-dose by hepatic function group (all hepatic function groups on the same plot per scale, based on the summary of concentrations by hepatic function group and time post-dose). - mean concentrations time plots (on both linear and semi-log scales) against nominal time post-dose by hepatic function group (all hepatic function groups on the same plot per pfizer general businessscale, based on the summary of concentrations by hepatic function group and time post-dose). - individual concentration time plots by hepatic function group (on both linear and semi-log scales) against actual time post-dose (there will be separate spaghetti plots for each hepatic function group per scale). for summary statistics, median and mean plots by sampling time, the nominal pk sampling time will be used, for individual subject plots by time, the actual pk sampling time will be used. a one-way analysis of variance (anova) will be used to compare the natural log transformed pf-06835919 aucinf, cci cmax, unbound aucinf (aucinf,u), cci and unbound cmax (cmax,u), as data permit, for each of the hepatic impairment cohorts (test) to the cohort without hepatic impairment (reference). estimates of the adjusted mean differences (test - reference), and corresponding 90% confidence intervals, will be obtained from the model. these will be exponentiated to provide estimates of the ratio of adjusted geometric means (test/reference) and 90% confidence intervals for the ratios. additionally, as a sensitivity analysis, age, gender and body weight will be explored as an additional covariate/factor in the model, as appropriate. linear regression will be used to analyse the potential relationship between appropriate natural log transformed pk parameters [aucinf, aucinf,u, cci] and hepatic function (serum albumin concentration, prothrombin time and total bilirubin). plots of pk parameters (eg, aucinf, aucinf,u, cci) versus hepatic function will be constructed, with a regression line and 90% confidence region included. estimates of the slope and intercept, together with a 90% confidence interval (ci), and the coefficient of determination (i.e. r-squared and adj-r-squared) will be obtained from the model. residuals from the models will be examined for normality and the presence of outliers via visual inspection of plots of residuals vs predicted values and normal probability plots of residuals but these will not be included in the clinical study report. if there are major deviations from normality or outliers then the effect of these on the conclusions will be investigated through alternative transformations and/or analyses excluding outliers. justification for any alternative to the planned analysis will be given in the report of the study.

Update

Type	Number	Title	Analysis Population	Sourc...	Mock Refere...	Similarity
☐ T	14.2.2.1	Descriptive Summary of [MATRIX] [ANALYTE] PK Parameters	PK Parameter Set	ADPP	PP Excel Mock S...	0.745396614074707
☐ T	14.2.2.2.1	Statistical Summary of [MATRIX] [ANALYTE] PK Parameters	PK Parameter Set	ADPP		0.5941721796989441
☒ F	14.2.2.2.2	Box and Whisker Plot of Individual and Geometric Mean for [MATRIX] [ANALYTE] [PAR...	PK Parameter Set	ADPP		0.51570725440979
☐ L	16.2.5.3.1	Listing and Descriptive Summary of [MATRIX] [ANALYTE] PK Parameters	PK Parameter Set	ADPP		0.795169472694397
☐ L	16.2.5.3.2	Supporting Data for Estimation of Thalf for [MATRIX] [ANALYTE]	PK Parameter Set	ADPP		0.1566862165927887

Custom NER

Look and Identify



autoLOT: An Illustrative Showcase

Analysis of specific domains

6.2.2 Demographic and Clinical Examination Data

SAP text

a summary of demographic data will be provided for age, race, weight, body mass index, and height. each will be summarized by hepatic function group and 'total' in accordance with the sponsor reporting standards.

Update

T...	Nu...	Title	Analysis Population	Sou...	Mock Referen...	Similarity
☐ T	14.1.2.1.x	Demographic Characteristics - Full Analysis Set	Full Analysis Set	ADSL	DM Excel Mock S...	0.4448639452457428
☐ T	14.1.2.1	Demographic and Baseline Characteristics - Full Analysis Set	Full Analysis Set	ADSL	DM Excel Mock S...	0.2462133765220642

Filters

Columns

Topic Modelling



autoLOT: An Illustrative Showcase

6.2.6. Vital Signs Data

SAP text

the **baseline measurement** (for blood pressure and pulse rate) is the last pre-dose measurement on day 1. absolute values and changes from baseline in seated systolic and diastolic blood pressure, and pulse rate will be summarized by hepatic function group and visit using sponsor reporting standards: **maximum absolute values and changes from baseline** or seated vital signs will be summarized descriptively by hepatic function group using categories as defined in appendix 2. numbers and percentages of participants meeting the categorical criteria will be provided. all planned and unplanned post-dose time-points will be counted in these categorical summaries. all values meeting the criteria of potential clinical concern will be listed. these data will be listed in accordance with the sponsor reporting standards. pfizer general business

Update						
Type	Nu...	Title	Analysis...	Sourc...	Mock Reference	Similarity
☐ T	14.3.8.x	Vital Signs: Results - Safety Set	Safety Set	ADVS	VS Excel Mock Shells	0.9268184304237366
☐ T	14.3.8.x	Vital Signs: Baseline and Change from Baseline - Safety Set	Safety Set	ADVS	VS Excel Mock Shells	0.9087928533554077
☐ T	14.3.8.x	Vital Signs: Maximum Change from Baseline - Safety Set	Safety Set	ADVS	VS Excel Mock Shells	0.6720688343048096
☐ T	14.3.8.x	Categorization of Vital Sign Results <Treatment Period>: Absolute Value and Maximum Change From ...	Safety Set	ADVS	VS Excel Mock Shells	0.5143527984619141
☐ L	16.2.8....	Vital Signs Data	Safety Set	ADVS	VS Excel Mock Shells	0.46240973472595215
☐ L	16.2.8....	Vital Signs Values in Specific Categories	Safety Set	ADVS	VS Excel Mock Shells	0.4142220914363861

1 to 6 of 6 |< < Page 1 of 1 > >|

Text
Similarity



autoLOT: An Illustrative Showcase

Consolidating all the required TLFs

☒ Selected TLF

T	Number	Title	Analysis Population	Source Datasets	Mock Reference
T	14.1.2.1.x	Demographic Characteristics - Full Analysis Set	Full Analysis Set	ADSL	DM Excel Mock Shells
T	14.1.2.1	Demographic and Baseline Characteristics - Full Analysis Set	Full Analysis Set	ADSL	DM Excel Mock Shells
T	14.2.2.1	Descriptive Summary of [MATRIX] [ANALYTE] PK Parameters	PK Parameter Set	ADPP	PP Excel Mock Shells
T	14.2.2.2.1	Statistical Summary of [MATRIX] [ANALYTE] PK Parameters	PK Parameter Set	ADPP	PP Excel Mock Shells
F	14.2.2.2.2	Box and Whisker Plot of Individual and Geometric Mean for [MATRIX] [ANALYTE] [PAR...	PK Parameter Set	ADPP	PP Excel Mock Shells
L	16.2.5.3.1	Listing and Descriptive Summary of [MATRIX] [ANALYTE] PK Parameters	PK Parameter Set	ADPP	PP Excel Mock Shells
T	14.3.8.x	Vital Signs: Results - Safety Set	Safety Set	ADVS	VS Excel Mock Shells
T	14.3.8.x	Vital Signs: Baseline and Change from Baseline - Safety Set	Safety Set	ADVS	VS Excel Mock Shells
T	14.3.8.x	Vital Signs: Maximum Change from Baseline - Safety Set	Safety Set	ADVS	VS Excel Mock Shells
T	14.3.8.x	Categorization of Vital Sign Results <Treatment Period>: Absolute Value and Maximu...	Safety Set	ADVS	VS Excel Mock Shells
L	16.2.8.2.x	Vital Signs Data	Safety Set	ADVS	PP Excel Mock Shells



Challenges, Best Practices & the Road Ahead

CHALLENGES

BEST PRACTICES

ROAD AHEAD

Non-standardized SAP input

Analyzing patterns, and improving efficiency



Enhancing SAP data quality by standardizing input

Building the repository

Construct and update the repository with completed studies



Maintain an up-to-date repository by constructing it with completed studies and regularly updating its contents.

Constructing a training dataset from scratch

Breaking down the elements and constructing organized information



Keep enhancing the dataset

Building custom models

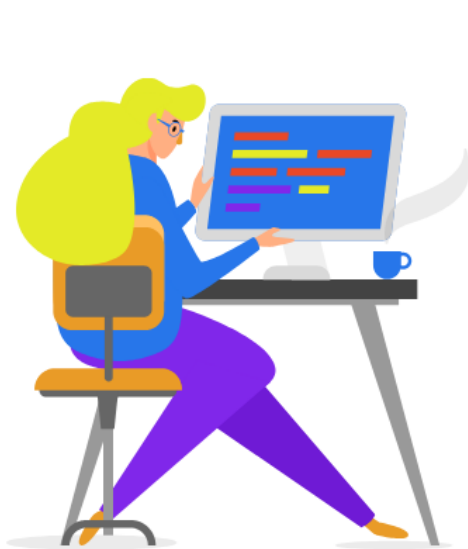
Mastering the intricacies of the human brain to teach machines to think alike



Refine the models through iterative process



Significant Takeaway



01

Accelerate the TLF generation process

02

Optimally utilize time and resources

03

Establish and manage a consistent repository throughout the entire process

Acknowledgement

- ❖ Debjit Biswas
- ❖ Priya Venkatesan Iyer
- ❖ Jeba Kumar
- ❖ Periasamy Karuppannan
- ❖ Aishwarya L R
- ❖ Abhinav Srivastava
- ❖ Francis Kevin Raj



**The Global Healthcare
Data Science Community**

Contact Channels

- 📞 UK +44 1843 609600
- ✉ office@phuse.global
- 🌐 phuse.global

Social Media

- 🐦 [@phusetwitta](https://twitter.com/phusetwitta)
- 📘 [/phusebook](https://facebook.com/phusebook)
- 📺 [/phusetube](https://youtube.com/phusetube)
- 🌐 [/company/phuse](https://linkedin.com/company/phuse)

An Overview of Natural Language Processing and Deep Learning

- Deep Learning is a branch of Artificial Intelligence (AI) that focuses on training computers to acquire knowledge from extensive datasets in order to carry out specific tasks. By utilizing neural networks, Deep Learning emulates the cognitive capabilities of the human brain, enabling the identification of meaningful patterns for making informed decisions.
- NLP, a branch of linguistics and AI, is dedicated to comprehending all aspects of human language. The objective of NLP tasks extends beyond the mere comprehension of individual words, aiming to grasp the contextual meaning behind them.

