Testing Prompt Generalization Robustness with ChatGpt

Once a given prompt has demonstrated it performs well producing content that matches the request, will the format of that prompt continue to perform well as it is generalized to handle similar conditions?

Introduction

When writing prompts, accounts of poor performance on task are often met with suggestions to changing the prompting in a ways that make it easier for the LLM to generate the desired response. Suggestions involve things such as "provide more context", "write exact steps", "ask it to think about its answer", "tell it to write the code and then produce the answer", and even the more polite "say please". There is a kind of implied absoluteness to such suggestions, as if the suggestion will always work, regardless the context. If a given task is described properly, with the proper context, that suggests such a task might be presented as a prompt template, allowing the insertion of variables into the prompt, that would produce similarly successful output.

Do prompts generalize, or is the definition of "the right prompt strategy" for every single question? If so, is there a way to find that "right prompt", or is "right" purely random, and our application of strategy just a way to motivate us to change the prompt. Roll the dice for another try at luck?

In this study, I performed a search for a generalizable task prompt for a problem that LLMs are supposed to be good at: extracting data from loosely structured content.

Summary of Findings

Hypothesis:

A prompt template that has been demonstrated to accurately extract and describe the input fields from a given form-based web page will perform with similar accuracy on other form-based web pages.

Conclusion:

The hypothesis has been demonstrated to be false. Using the same template, and changing only the name of the application and the web page it referred to, a prompt template that had resulted in 100% correct extraction on one site demonstrated wide range of performance – from 100% on, to around 80% and even as low as -114% (meaning it got more wrong fictional guess than there were items to extract from the page).

A prompting strategy that has been demonstrated to perform better on a specific website for the task described above is insufficient to generalize the same task across other websites of similar form and structure. Prior good performance is insufficient, meaning if there is a well-performing

generalized template for this task, some missing attribute of the prompt is required that was not utilized in this experiment.
Another possibility is that there is no such thing as a robust, high performing template which generalizes against a wide variety of inputs. It is possible that the inputs themselves (application name, URL – or the content of that site as represented in the training data) are sufficient to affect performance, skewing it better or worse.

Summary of Outcomes

Site	Actual	Extracted	Extra	Score (Extracted – Extra)/Actual
YouTube Notification Settings	14	5	3	14%
DeviantArt Sign-up	5	4		80%
PayPal Business Account Sign-up	10	10		100%
Washington State Business License Wizard	14	2	18	-114%
OSHA ITA Coverage Application	5	5		100%

Challenges to the Outcomes

Small sample size

The data set was small. The hypothesis was not framed as how well the prompting strategy performed, but whether the LLM could be demonstrated to fail (less than 100% on accurate extraction) on similar sites without changing the prompt template. Were the hypothesis about performance of the strategy, a much larger sample size would be needed.

Insufficiently selected/crafted prompt

It is possible that the initial selection yielded a prompt that scored 100% by luck rather than prompt robustness. This is a credible challenge to the approach that deserves follow-up investigation. One way of doing that might be to take the same set of sets and iterate on modifications to the prompt template until all those sites yield 100% success. For that purpose, a larger data set would be advised.

Temperature setting was not adjusted

This trial was done with the web-client for ChatGpt, with the temperature settings left at default (non-zero). Randomness in token prediction may have contributed to failures to produce accurate output.

Site data being in training data or not

It is possible not all the sites in the test were in the training data, or that the version of the site ChatGpt was trained on is different than the site content at the time of training. This is a valid challenge to the methodology and deserves further examination. Some means of doing so might involve some means of checking whether the site is in the training data or not, or extracting the actual form content and passing it through the prompt.

Site variations perturbing results

One of the sites (Washington State Business License Wizard) was a multi-step form rather than a one-page form. It is possible those differences might have accounted for differences in performance on that site in a way that would have necessitated changing the prompt to account for it. It is possible other non-obvious differences in site behavior might account for other variations in performance, although in the latter case those would suggest support for the hypothesis being disproved, as robustness across different sites is part of prompt generalization to this task.

Methodology

Prompt Selection and Generalization

1. Prompt ChatGpt to extract the fields from a web-form
2. Check the result for correctly matched, missing, and extra fields – scoring = (matched – extra)/actual
3. With new sessions each time, alter the prompt to attempt higher accuracy
4. Once one of the prompts achieves 100% accuracy (no missing fields, no extra fields), generalize that prompt

Trial prompts:

Attempt #1: (had a missing field)

Attempt #1 is a two step approach, URL given first with a check if ChatGpt is aware of the page. The results produced exact matching of precise words that suggest the page was in the training data.

> *Are you able to see this page? https://www.osha.gov/itareportapp*
>
> ***ChatGPT said:***
>
> *I can't access web pages directly, but you can describe what you're looking for on that OSHA page, and I'll do my best to help!*
>
> *4o*
>
> ***You said:***
>
> *I am trying to identify a test matrix for that page based on the input values of the form on the page. Please generate a list of the input variables for that form. Just the variables and their data type.*

Attempt #2: (generated extra fields)

For Attempt #2, the application name and URL were put into an attempt at a one-shot request.

> *The ITA Coverage Application OSHA website has a form (https://www.osha.gov/itareportapp). Construct a list of the relevant variables one would use as inputs to the form for testing purposes, along with a list of their data types.*

Attempt #3: (100% correct)

For Attempt #3, the prompt is nearly identical with the addition of the phrase "*based on the input fields for the page*". This yielded a 100% match, so this prompt was selected for generalization.

> *The ITA Coverage Application OSHA website has a form (https://www.osha.gov/itareportapp). Construct a list of the relevant variables one would use as inputs to the form for testing purposes based on the input fields for the page, along with a list of their data types*

Selected Prompt:

The selected prompt was generalized by replacing the site application name ("*ITA Coverage Application OSHA"*) and the URL ("*https://www.osha.gov/itareportapp*") with variables that were substituted with alternate applications and URLs for the trials.

```
The <application name> website has a form (<URL>). Construct a list of
the relevant variables one would use as inputs to the form for testing
purposes based on the input fields for the page, along with a list of
their data types
```

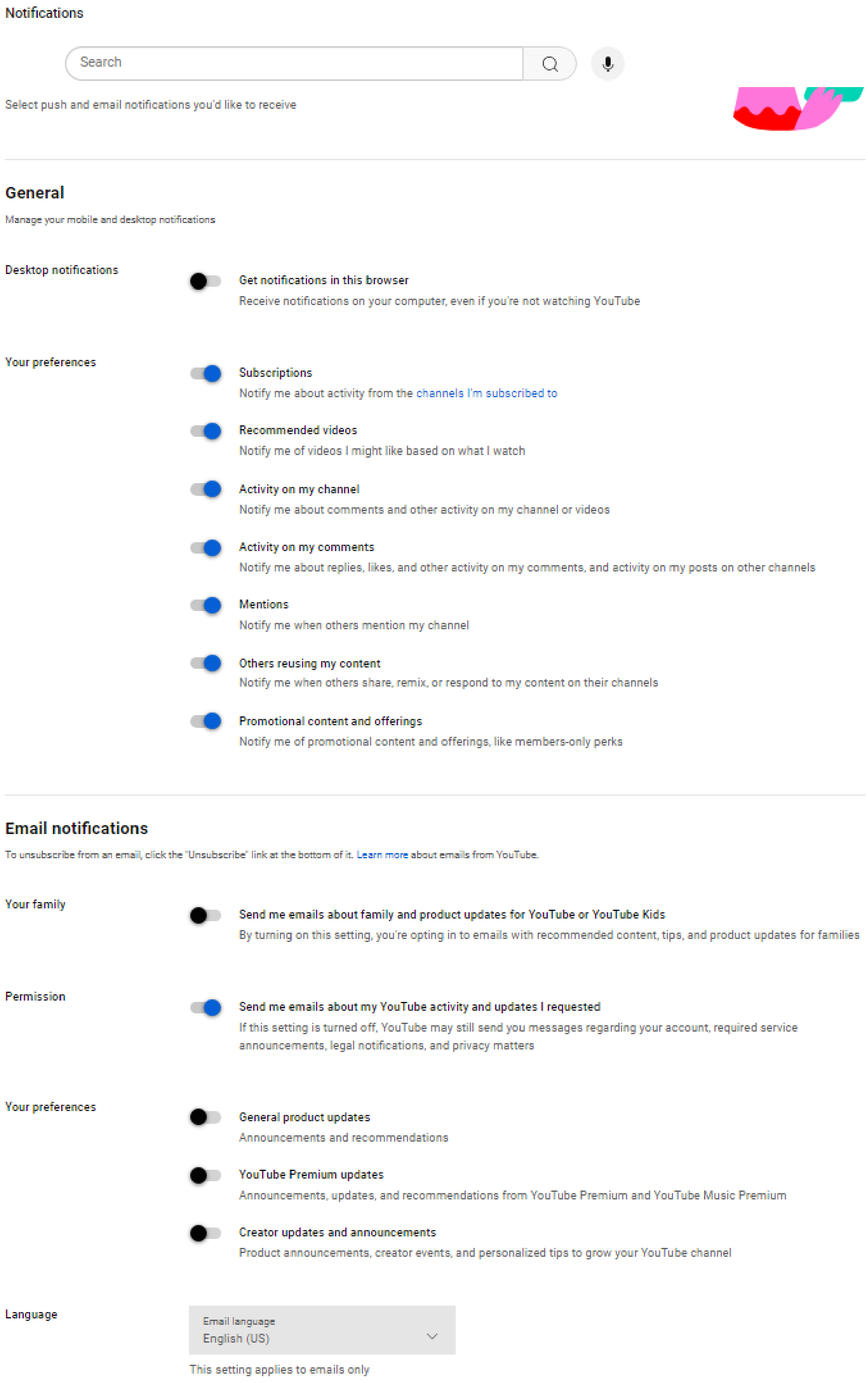
Assessment Methodology:

1. Replace **<application name>** and **<URL>** with the names of different web-based form applications. Execute the prompt in ChatGpt
2. Compare the identified input variables in the response with a human generated list of the input variables for each site.
3. Scoring:
    a. Each matched variable = +1
    b. Each missed variable = 0
    c. Each extra variable or incorrectly identified (e.g. type is wrong) = -1
    d. Actual number of variables by total to derive correctness as a percentage (value can be negative)

Appendix

YouTube Notifications Settings

https://chatgpt.com/share/67c3406e-c534-800a-a2ac-af6db81df010



Notifications

Select push and email notifications you'd like to receive

General

Manage your mobile and desktop notifications

Desktop notifications

○ Get notifications in this browser
Receive notifications on your computer, even if you're not watching YouTube

Your preferences

● Subscriptions
Notify me about activity from the channels I'm subscribed to

● Recommended videos
Notify me of videos I might like based on what I watch

● Activity on my channel
Notify me about comments and other activity on my channel or videos

● Activity on my comments
Notify me about replies, likes, and other activity on my comments, and activity on my posts on other channels

● Mentions
Notify me when others mention my channel

● Others reusing my content
Notify me when others share, remix, or respond to my content on their channels

● Promotional content and offerings
Notify me of promotional content and offerings, like members-only perks

Email notifications

To unsubscribe from an email, click the "Unsubscribe" link at the bottom of it. Learn more about emails from YouTube.

Your family

○ Send me emails about family and product updates for YouTube or YouTube Kids
By turning on this setting, you're opting in to emails with recommended content, tips, and product updates for families

Permission

● Send me emails about my YouTube activity and updates I requested
If this setting is turned off, YouTube may still send you messages regarding your account, required service announcements, legal notifications, and privacy matters

Your preferences

○ General product updates
Announcements and recommendations

○ YouTube Premium updates
Announcements, updates, and recommendations from YouTube Premium and YouTube Music Premium

○ Creator updates and announcements
Product announcements, creator events, and personalized tips to grow your YouTube channel

Language

Email language
English (US)

This setting applies to emails only

Actual	Extracted	
Get Notifications in Browser		0
Subscriptions		0
Recommended Videos	Recommended Videos	+1
Activity on my Channel	Activity on your channel	+1
Activity on my Comments	Comment Notifications	+1
Mentions	Mentions	+1
Others reusing my comments		0
Promotional Content and Offerings		0
Send me emails about family and product updates		0
Send me emails about my YouTube activity and updates		0
General Product Updates	Product Updates	+1
YouTube Premium Updates		0
Creator Updates and announcements		0
Language (dropdown list)		0
	Email Notifications	-1
	Push Notifications	-1
	Subscription Notifications (improperly classified as a dropdown with three options)	-1

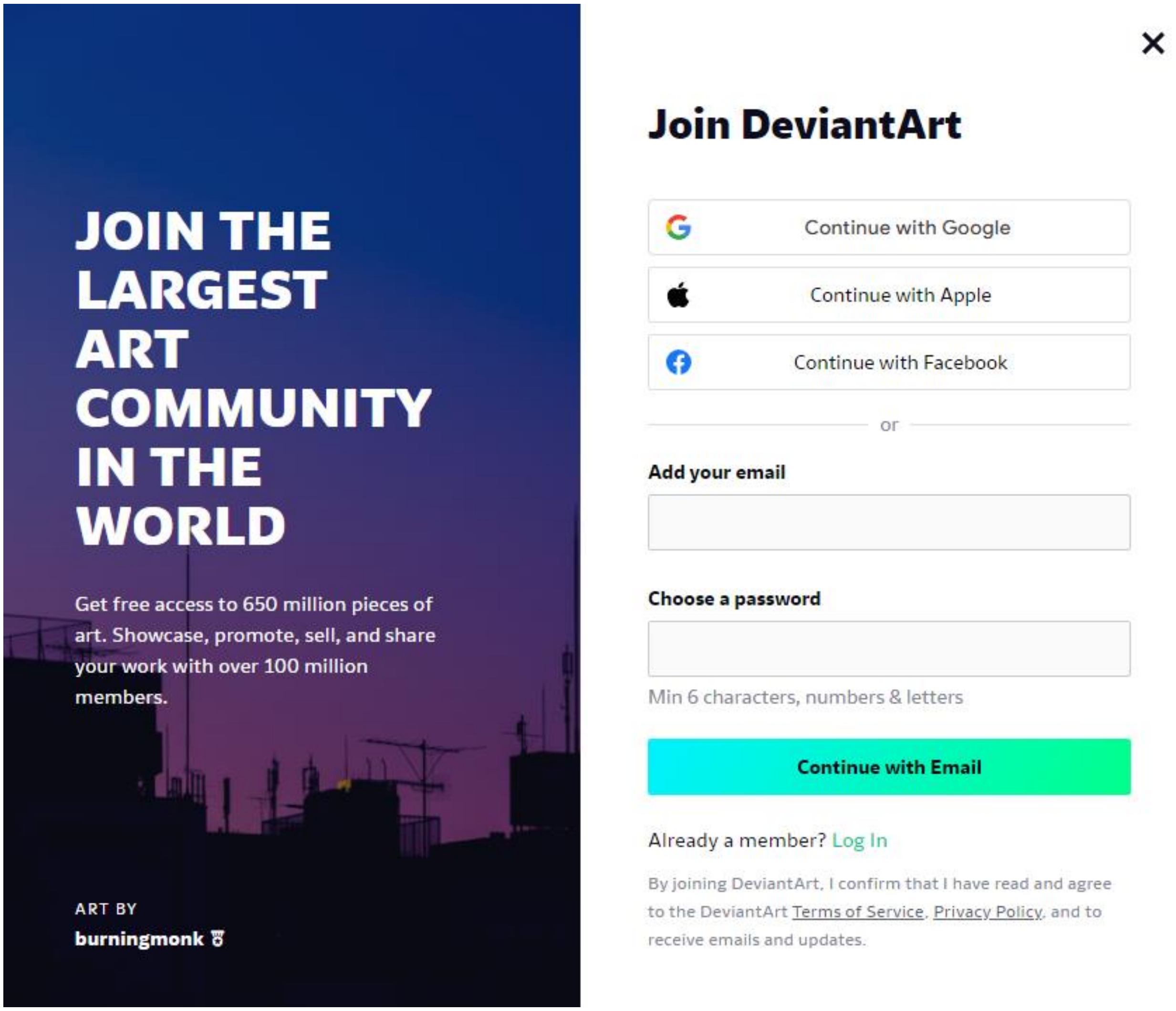
5 correct

3 extra

(5-3)/14=14%

DeviantArt Sign-up Page

https://chatgpt.com/share/67c33ebc-3b78-800a-b39d-beee8e2214d4



JOIN THE LARGEST ART COMMUNITY IN THE WORLD

Get free access to 650 million pieces of art. Showcase, promote, sell, and share your work with over 100 million members.

ART BY
burningmonk

Join DeviantArt

Already a member? Log In

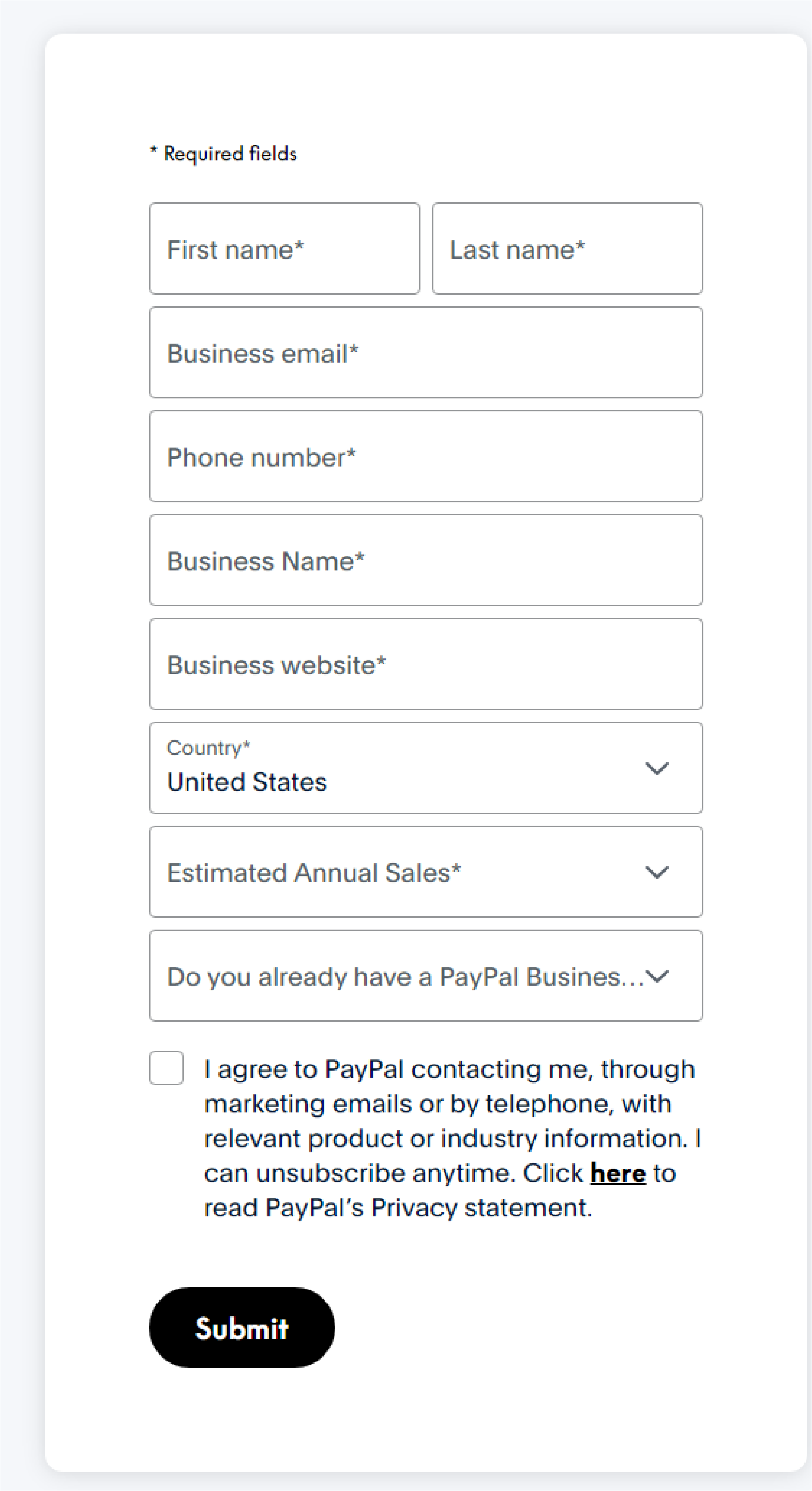
By joining DeviantArt, I confirm that I have read and agree to the DeviantArt Terms of Service, Privacy Policy, and to receive emails and updates.

Actual	Extracted	
Continue with Google		0
Continue with Apple	Apple Id	+1
Continue with Facebook	Facebook	+1
Email	Email	+1
Password	password	+1

= 4/5=80%

PayPal Business Account Signup

https://chatgpt.com/share/67c33e1a-9b84-800a-981d-51f557872450



* Required fields

First name*

Last name*

Business email*

Phone number*

Business Name*

Business website*

Country*
United States

Estimated Annual Sales*

Do you already have a PayPal Busines…

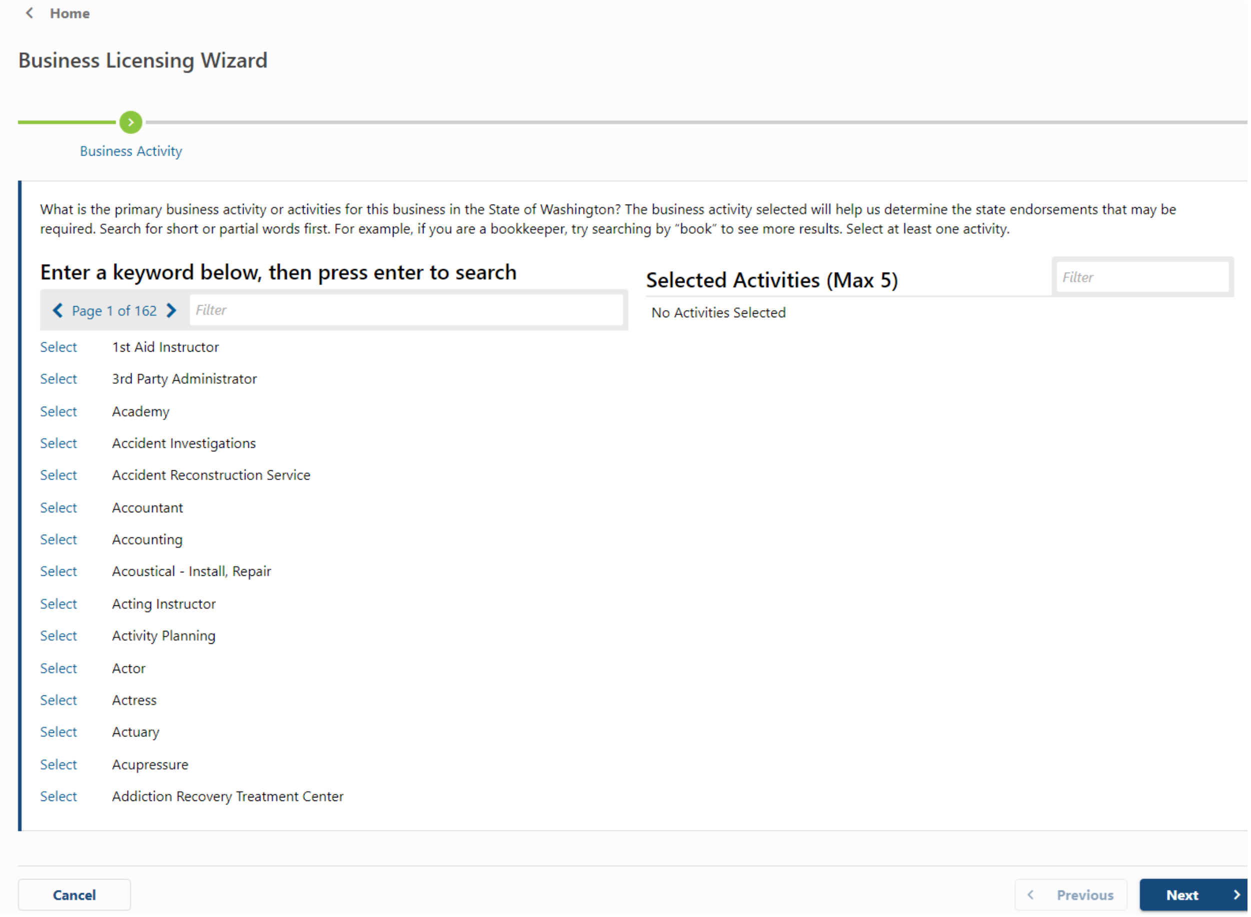
□ I agree to PayPal contacting me, through marketing emails or by telephone, with relevant product or industry information. I can unsubscribe anytime. Click **here** to read PayPal's Privacy statement.

Actual	Extracted	
First	First name	1
Last	Last name	1
Business Email	Business email	1
Phone Number	Phone number	1
Business Name	Business Name	1
Business Website	Business Website	1
Country	Country	1
Estimated Annual Sales	Estimated Annual Sales	1
Do you already have a PayPal Business...	Do you already have a PayPal Business Account	1
Agree to contact	Agree to contact	1

10/10=100%

Washington State Business License Wizard

https://chatgpt.com/share/67c33bae-54d8-800a-a033-a229096e6917



Business Licensing Wizard

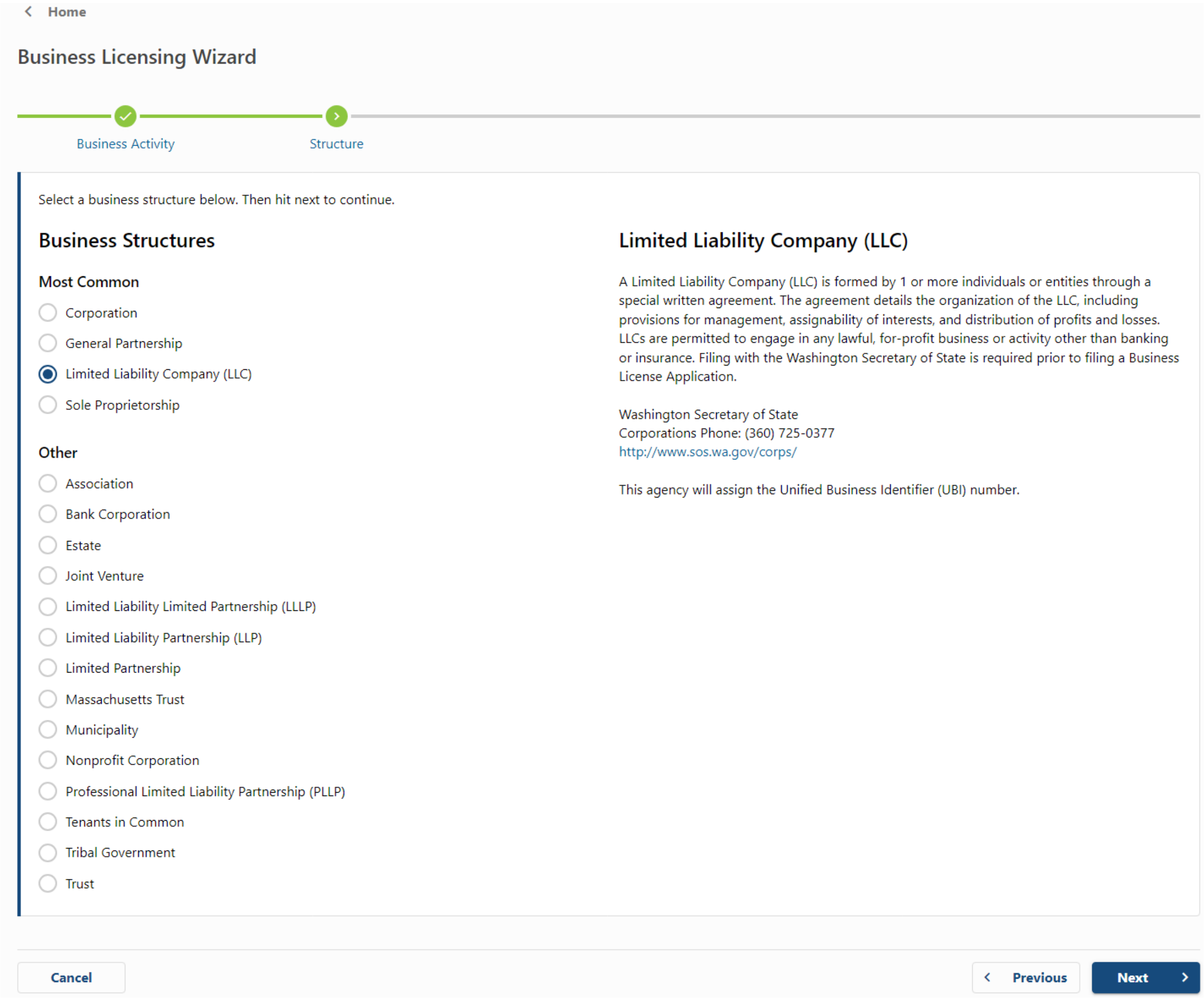
What is the primary business activity or activities for this business in the State of Washington? The business activity selected will help us determine the state endorsements that may be required. Search for short or partial words first. For example, if you are a bookkeeper, try searching by "book" to see more results. Select at least one activity.

Enter a keyword below, then press enter to search

Selected Activities (Max 5)

No Activities Selected

Select	1st Aid Instructor
Select	3rd Party Administrator
Select	Academy
Select	Accident Investigations
Select	Accident Reconstruction Service
Select	Accountant
Select	Accounting
Select	Acoustical - Install, Repair
Select	Acting Instructor
Select	Activity Planning
Select	Actor
Select	Actress
Select	Actuary
Select	Acupressure
Select	Addiction Recovery Treatment Center



Select a business structure below. Then hit next to continue.

Business Structures

Most Common

○ Corporation
○ General Partnership
● Limited Liability Company (LLC)
○ Sole Proprietorship

Other

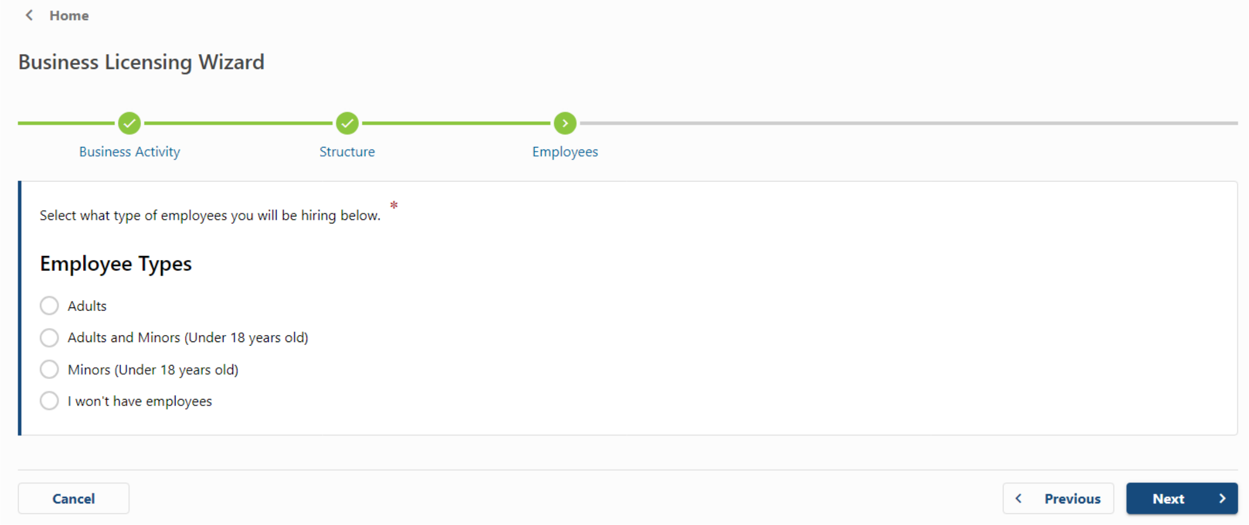
○ Association
○ Bank Corporation
○ Estate
○ Joint Venture
○ Limited Liability Limited Partnership (LLLP)
○ Limited Liability Partnership (LLP)
○ Limited Partnership
○ Massachusetts Trust
○ Municipality
○ Nonprofit Corporation
○ Professional Limited Liability Partnership (PLLP)
○ Tenants in Common
○ Tribal Government
○ Trust

Limited Liability Company (LLC)

A Limited Liability Company (LLC) is formed by 1 or more individuals or entities through a special written agreement. The agreement details the organization of the LLC, including provisions for management, assignability of interests, and distribution of profits and losses. LLCs are permitted to engage in any lawful, for-profit business or activity other than banking or insurance. Filing with the Washington Secretary of State is required prior to filing a Business License Application.

Washington Secretary of State
Corporations Phone: (360) 725-0377
http://www.sos.wa.gov/corps/

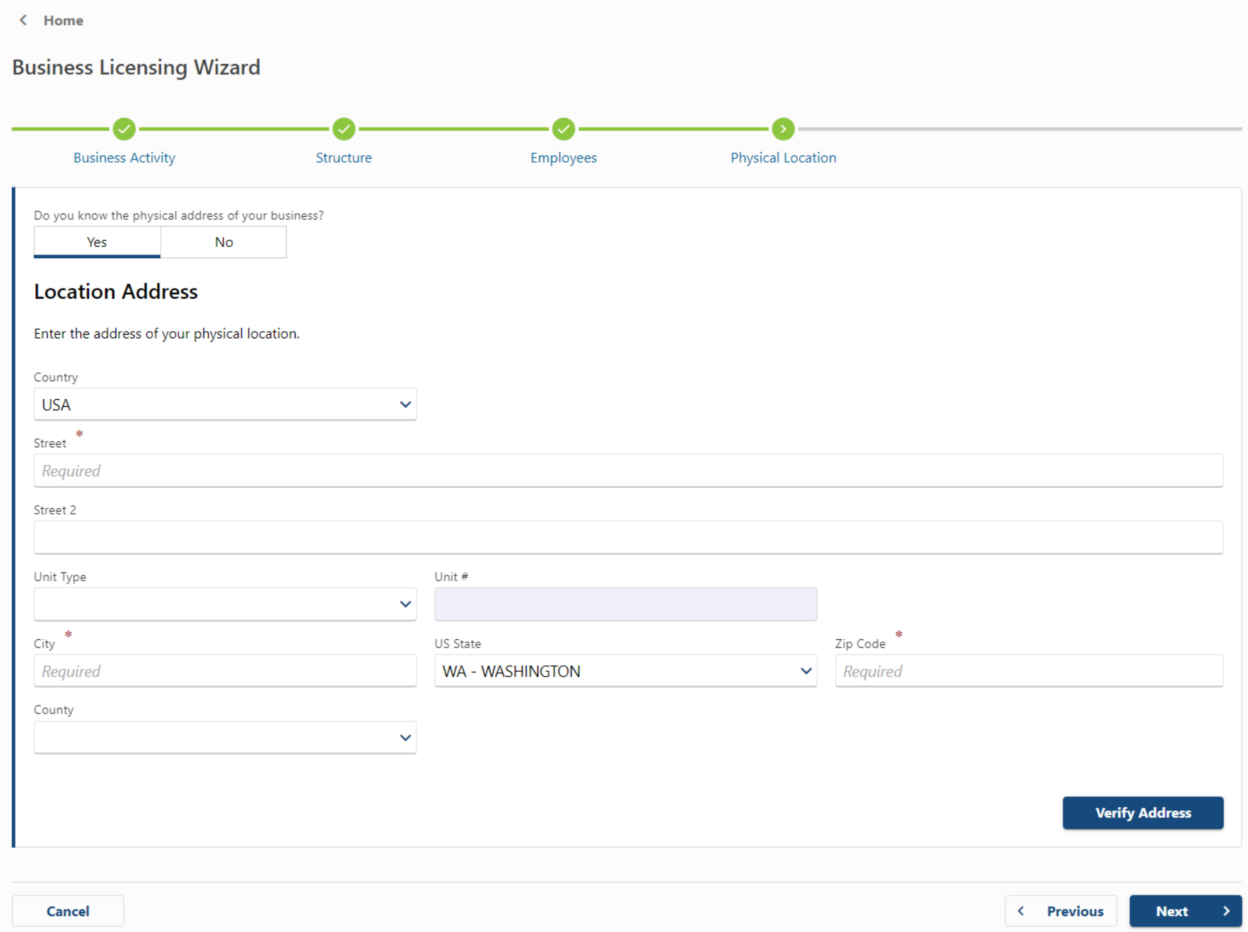
This agency will assign the Unified Business Identifier (UBI) number.



Select what type of employees you will be hiring below. *

Employee Types

○ Adults
○ Adults and Minors (Under 18 years old)
○ Minors (Under 18 years old)
○ I won't have employees



Do you know the physical address of your business?

Yes | No

Location Address

Enter the address of your physical location.

Country: USA

Street *: Required

Street 2:

Unit Type:

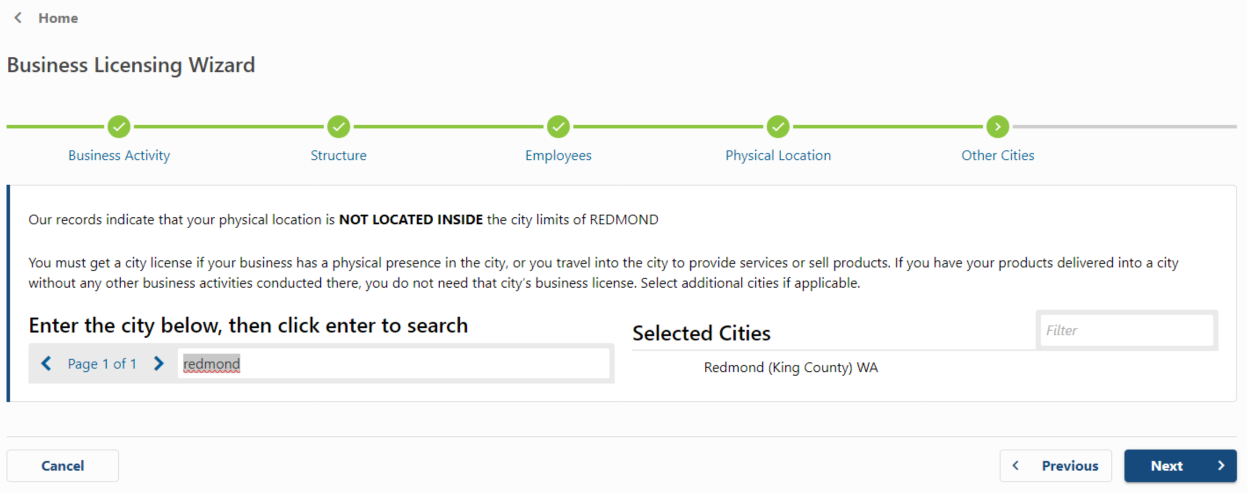
Unit #:

City *: Required

US State: WA - WASHINGTON

Zip Code *: Required

County:



Our records indicate that your physical location is **NOT LOCATED INSIDE** the city limits of REDMOND

You must get a city license if your business has a physical presence in the city, or you travel into the city to provide services or sell products. If you have your products delivered into a city without any other business activities conducted there, you do not need that city's business license. Select additional cities if applicable.

Enter the city below, then click enter to search

‹ Page 1 of 1 › redmond

Selected Cities

Filter

Redmond (King County) WA

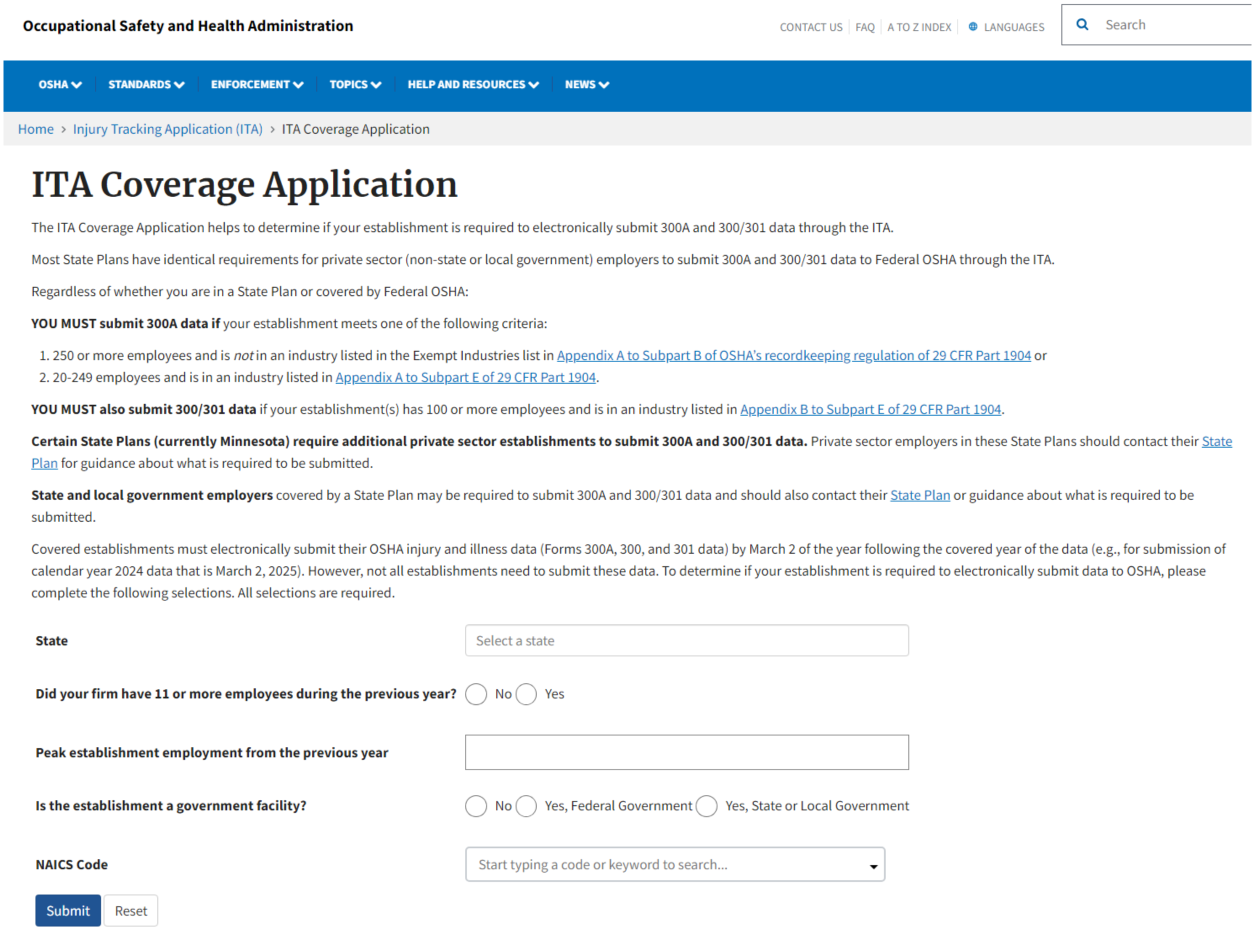
Actual	Extracted	
Activities	Description of Activities	1
Business Structure (pre-defined list)	Business Structure	1
Other (pre-defined list)		0

Employee Types (pre-defined list)		0
Country		0
Street 1		0
Street 2		0
Unit Type		0
Unit #		0
City		0
State		0
Zip		0
County		0
City of registration		0
	Business Name	-1
	Owner Full Legal Name	-1
	Unified Business Identifier	-1
	Federal Employee Identification Number	-1
	Business Address	-1
	Mailing Address	-1
	Business Phone Number	-1
	Business Email Address	-1
	NAICS Code	-1
	Date Business will Start	-1
	Number of Employees	-1
	Employment Security Department Number	-1
	Department of Labor & Industries (L&I) Account ID	-1
	Gross Annual Revenue Estimate	-1
	Professional Licenses Held	-1
	Trade Name/DBA	-1
	Registered Agent Information	-1
	Previous Business Name	-1

(2 – 18) / 14 = -114%

OSHA ITA Coverage Application Form

https://chatgpt.com/share/67c334f6-8ddc-800a-9ebf-68d7b22dc1e4


Occupational Safety and Health Administration
CONTACT US | FAQ | A TO Z INDEX | LANGUAGES
Search
OSHA
STANDARDS
ENFORCEMENT
TOPICS
HELP AND RESOURCES
NEWS
Home › Injury Tracking Application (ITA) › ITA Coverage Application

ITA Coverage Application

The ITA Coverage Application helps to determine if your establishment is required to electronically submit 300A and 300/301 data through the ITA.

Most State Plans have identical requirements for private sector (non-state or local government) employers to submit 300A and 300/301 data to Federal OSHA through the ITA.

Regardless of whether you are in a State Plan or covered by Federal OSHA:

YOU MUST submit 300A data if your establishment meets one of the following criteria:

1. 250 or more employees and is *not* in an industry listed in the Exempt Industries list in Appendix A to Subpart B of OSHA's recordkeeping regulation of 29 CFR Part 1904 or
2. 20-249 employees and is in an industry listed in Appendix A to Subpart E of 29 CFR Part 1904.

YOU MUST also submit 300/301 data if your establishment(s) has 100 or more employees and is in an industry listed in Appendix B to Subpart E of 29 CFR Part 1904.

Certain State Plans (currently Minnesota) require additional private sector establishments to submit 300A and 300/301 data. Private sector employers in these State Plans should contact their State Plan for guidance about what is required to be submitted.

State and local government employers covered by a State Plan may be required to submit 300A and 300/301 data and should also contact their State Plan or guidance about what is required to be submitted.

Covered establishments must electronically submit their OSHA injury and illness data (Forms 300A, 300, and 301 data) by March 2 of the year following the covered year of the data (e.g., for submission of calendar year 2024 data that is March 2, 2025). However, not all establishments need to submit these data. To determine if your establishment is required to electronically submit data to OSHA, please complete the following selections. All selections are required.

State Select a state

Did your firm have 11 or more employees during the previous year? ○ No ○ Yes

Peak establishment employment from the previous year

Is the establishment a government facility? ○ No ○ Yes, Federal Government ○ Yes, State or Local Government

NAICS Code Start typing a code or keyword to search...

Actual	Extracted	
State	State	1
Does your firm have > 11 people	Does your firm have > 11 people	1
Peak establishment employment	Peak establishment employment	1
Is the establishment a government facility	Is the establishment a government facility	1
NAISC Code	NAISC Code	1

5/5=100%